12/21/98 6:39 PM

# Reciprocal Fairness and Social Signaling: Experiments with Limited Reputations.

Catherine C. Eckel
Department of Economics
Virginia Polytechnic Institute and State University
Blacksburg, VA  24061
eckelc@vt.edu

and

Rick K. Wilson
Department of Political Science
Rice University
Houston, TX 77251-1892
rkw@rice.edu

This paper was prepared for presentation at the American Economic Association Meeting, January 3-5, 1998, New York, NY.   The authors thank Sean Bolks and Laura Degado for help in running these experiments.

“Just put on a Happy Face!”
-- Lee Adams and Charlie Strouse

## Introduction

Initial impressions are important for building reputation. While an initial impression does not constitute a complete reputation, it is critical for determining how someone is judged, and how subsequent information about that person is interpreted. The voluminous literature on stereotyping confirms this point (see MacRae, et al., 1996, for a survey.) While the point may seem obvious, it has not been an important element of game-theoretic models of human strategic behavior.

Reputation is especially important when modeling repeated games (see for instance Kreps et al. 1982, and Fudenberg and Maskin, 1986; Fudenberg and Levine, 1992). Such models typically assume that actors have an initial set of expectations about their counterpart’s strategies. The play of the game, whether repeated or sequential, allows an actor to update those priors and then rely on the counterpart’s reputation. In repeated games, the ability to both infer and develop a reputation can be critical for reaching a Pareto-superior equilibrium. However, the source of initial reputation is a matter that is largely unexplored. For single-shot games and the initial play of repeated games, reputational priors are critical. For games with multiple equilibria, the equilibrium that is reached is often characterized as path-dependent, and the initial expectations formed by partners prior to playing the game are important for defining which path is taken.

We begin by observing that most exchange involves social interaction. In environments as different as negotiating a trade or voting for a candidate, an actor must decide whom to trust, select among alternative partners, assess the character of a partner, and choose a strategy based on that assessment. It is arresting to note that given incomplete information about a partner, people are quite good at gauging how to deal with a particular partner, and at the same time, are very accomplished at sending signals to others about their own intentions. Frank, et al., (1993), for example, show that subjects are able to predict the strategy choices of others when given the opportunity to observe them before the decision takes place. This ability to "read" one another makes the mix of behavioral strategies much richer than many game-theoretic models predict.

Our concern is with *social signaling*, in which agents seek cues about others' intentions, and send cues about their own. These cues build expectations for the behavior of others. We focus on a specific set of cues: nonverbal facial expressions. Much of game theory assumes actors are anonymous and have no information about the identities of their partners. However, we think that most social exchange involves face-to-face interaction and laboratory experiments demonstrate that such settings are very different from anonymous interaction. We eliminate verbal communication and focus on a relatively simple "cheap talk" device – stylized facial expressions – to examine how initial priors are formed about an unfamiliar exchange partner.

We want to understand the process by which reputations are generated. The settings we investigate are simple, one-shot, two-person sequential games. Each game has a unique subgame perfect equilibrium. A subset of the games contain an outcome that Pareto-dominates the equilibrium, but can only be reached by a sequence of moves that depends on actors trusting one another and engaging in reciprocal altruism. Both trust and reciprocity are grounded in the priors that actors hold about their counterparts in the game; evolutionary psychology provides a basis for inferring the source of an actor's prior.

In the next section we sketch the motivation for concern about the consistency of out-of-equilibrium behavior. In the second section we review some of the extensive literature on facial expressions, largely drawn from social psychology. Because human faces can produce complex blends of expressions, we adopt simplified facial icons. We then present survey and experimental results that test the impact of facial expressions on expectations and behavior. The final section concludes with a discussion of avenues of future research and the importance of these findings for game theory.

## Motivation

Traditional game theory describes rational actors who play best-response to the strategies of others which yields a Nash Equilibrium. Consider, for example, two individuals who not only stumble upon one another, but also stumble upon a freshly baked blueberry pie. They agree that one will cut the pie and put whatever piece she wishes on her plate, leaving the remainder for the other. However, before the pie can be eaten, the second party has the right to throw both pieces on the ground and stomp them into the earth. The first player will leave the other a tiny slice, based on the logic of looking inward and asking: "what would I do if I was only offered a tiny slice of the pie? Well, a tiny slice is better than nothing, which is what will happen if both pieces are thrown away in disgust. Besides, I like blueberry pie and want as much as I can get." Neither party to this exchange needs to go beyond looking inwardly and making a decision on that basis.

However, suppose the first actor momentarily looks outward. While contemplating the division, she notices that the second actor is agitated, is darting glances between the pie and herself, and is stomping his feet. Perhaps his actions are little more than bluster. On the other hand he may be intending to destroy the pie if he does not get considerably more than a tiny sliver. Unfortunately for the first actor, she has no priors about him -- whether he has a reputation for getting angry over being left little. All she can do is to try to read his intentions.[1]

While game theoretic models dominate the way economists think about and predict behavior, a large number of empirical papers show that laboratory subjects often do not play best-response, and eschew playing even dominant strategies (where available). Many of these papers are centrally concerned with actors looking beyond themselves and to their partners. This other-regarding behavior occurs in may settings including ultimatum and dictator games (Camerer, 1998; Eckel and Grossman, 1996, 1998; Forsythe, et al., 1994; Hoffman, et al., 1994), public goods games (Ledyard, 1995), investment trust games (Berg et al., 1995) and gift exchange experiments (Fehr, et al., 1993). These results are neither random nor haphazard. Behavior inconsistent with game theoretic predictions is routine and patterned. Moreover, a variety of explanations has been proposed for why game theoretic models fare poorly in some environments.

One promising approach suggests that humans have an evolved capacity to read the intentions of others. In psychology the primary focus of research on reading the intentions of others has been with autistics and primates (see Baron-Cohen, 1995 and O'Connell, 1998). Referred to as the "theory of mind" (TOM), psychologists are interested in the capacity of individuals to put themselves in the place of others. That is, can I separate myself from what I know or understand and try to comprehend what another might know? A simplified test of this concept is often given to children. A familiar brand of candy is shown a child and the child is asked to guess its contents. The child ordinarily guesses that candy is in the box, but then is shown that it contains pencils. After expressing surprise, the child is told that another person will come into the room and will be asked what is in the box. The child is asked to predict what that person will say. Most children over the age of three believe that the response will be

[1] This tale is not out of the realm of possibilities. For an amusing anecdote of bluster, see David Remnick's biography of Mohammed Ali. Here he relates the story of when Ali (then Cassius Clay) first faced Sonny Liston, the fearsome heavyweight champion of the time. Prior to the fight, Ali was expected to lose the fight. But Ali worked hard to establish a prior reputation with Liston, and Remnick attributes his surprising victory in part to Ali's prior behavior. Ali behaved as if he was trying to convince Liston that he was crazy, pretending to go berserk at the weigh-in, lunging and screaming at Liston. He later admitted this was all an act to get under Liston's skin. (Interview with Remnick, NPR Morning Edition, November 6, 1998)

"candy"; they are able to place themselves into another's shoes, imagining what it is like to be the other. By contrast, autistic children invariably predict that others will think the box contains pencils. They are unable to disentangle what they know from what the other might know. In other words, they are unable to look beyond their own knowledge and understand the mental state of the other. The inability to pick up on the intentions of others leads to serious problems for social interaction. The problem is not one of social ineptness, but rather a failure of cognitive ability to imagine the mental state of another person.

In economics this point is echoed by Robert Frank (1997). He wonders why people would ever tip in a restaurant in a city in which they are only visiting (as well as other, seemingly non-rational behaviors). He contends that

> "In this scheme, *sympathy* plays as important a role as it did in Adam Smith's scheme. I am inclined to leave the waiter in an out-of-town restaurant a tip because I feel sympathetic to the waiter's interest. I imagine myself in the waiter's position, having worked hard to provide good service, and how distressed I would feel if somebody failed to tip me. Sympathy is one of the key emotions for supporting the kind of behavior I have talked about in the examples." (Frank, 1997, p. 290).

Central to Frank's argument is that it is critical to understand the mental state of the other and that this is missing in standard models of game theory. The same point has been noted in a number of laboratory experiments that have centered on the concept of "reciprocity" [Berg, et al. (1995), Falk and Fischbacher (1998), Fehr and Schmidt (1997), Güth (1995a), McCabe, Rassenti and Smith (1998)]. In instances where there are multiple equilibria, the problem becomes one of capturing gains to coordination on a pareto-superior equilibrium (see van Hyuck et al. 1990 and Wilson and Rhodes, 1997). In other instances there are gains to consistent out-of-equilibrium play. Norms of reciprocity can capture these gains by serving as an equilibrium-selection mechanism, or by coordinating out-of-equilibrium play.

McCabe, Rassenti and Smith (1998) draw on evolutionary psychology to propose that actors have evolved distinct mental modules for social exchange that serve as equilibrium selection mechanisms. Such mental modules are adaptive and enable individuals to solve problems of multiple equilibria or to settle on Pareto-enhancing outcomes. Others suggest that evolutionary pressures operate on cultures in which cultural norms for reciprocity evolve to solve the equilibrium selection problem (Bowles, 1998; Bowles and Gintis, 1998; Boyd and Richerson, 1985). Whichever is true, it is clear that the ability to recognize and draw inferences about the mental states of others is useful. Whether culturally bestowed or adaptively endowed, an agent's capacity to quickly assess another is an important component of social exchange. That capacity can lead to a very simple initial choice: trust and cooperate with "friends", and withhold trust and limit cooperation with those perceived as threatening or uncooperative. Whether mental modules operate or not, it seems plausible that individuals pay close attention to the social signals of others and that reciprocal altruism emerges in many exchange contexts.

This approach sidesteps the debate over whether particular preferences have been selected by evolutionary pressures (e.g., evolved altruistic preferences) or whether a culture of "reciprocators" or conditional cooperators has evolved. This approach argues that human cognitive capacities are modular and highly attuned to the context within which strategic play takes place. Consequently other-regarding behaviors are likely to be contingent. Rather than addressing whether particular preferences have been "selected for", we focus on the mechanism that might triggers a choice between two alternative strategies: one that is consistent with a traditional model of utility maximization, and another that requires trust and reciprocity.

*Games.*

To illustrate circumstances under which reading the intention of another comes into play, consider the following two games. Both are two-person sequential games with perfect and complete information. The first game we characterize as a simple Nash game while the second game also has a Nash equilibrium, but incorporates gains to out-of-equilibrium play.

First consider the game given in extensive form by Figure 1. Two actors, A and B, face a series of moves through the game. Payoffs are given as dollars. By convention Actor A's payoffs are in the upper position and B's are in the lower. Actor A is given the first move and can choose to end the game, with an outcome of 9 (and B receiving 29), or he can pass the move to B. If the move is passed, then B has a similar choice -- end the game with A receiving 19 and B receiving 14 or pass the move back to A who then chooses between equivalent outcomes (4 units each).

<Figure 1 About Here>

The outcome to this game is straightforward. Under backward induction, B knows that she will obtain 4 if she passes the move to A at the middle node (there is no strategic choice at the last node because all outcomes are equivalent for A). However, if B chooses to quit at the second node, then she will obtain 14 units. If the choice is passed to her at this node, then she will choose to quit. A then knows that he will receive 19 if he passes the move to B at the first node. On the other hand if A quits at the first node, then he will obtain 9 units. Clearly A will pass at the first node and B will choose to quit. The outcome at the second node constitutes the unique Nash equilibrium for the game.

Now consider the second game given by Figure 2. Using backward induction, the unique Nash equilibrium is for Actor A to quit at the first node. At the last choice node, A will choose to move right rather than down. Knowing this, B would choose to quit at the second node, where she is left better off than passing the move back to A. However, at that node A only gets 3 units. Therefore he will choose to quit at the first node. Notice that unlike the first game, both A and B could do better at the last node than the first node. However, this can only happen if both actors could trust one another and if that trust is reciprocated. Certainly Actor A can signal his trust in Actor B by choosing to pass. This is the centerpiece of models relying on forward induction. What does A have to know about B in order to be assured that B will reciprocate A's trust?

<Figure 2 About Here>

The first game sketched above requires that actors know nothing about one another. Even if they understood one another's intentions, there is no advantage to doing anything except selecting the Nash equilibrium. For the second game, if trusting and reciprocal behavior is possible, both actors do better than in equilibrium. But in order to trust, Actor A has to make some inference about the trustworthiness of B. This goes to the heart of reading intentionality. Most approaches to understanding trustworthy behavior have started by thinking about how reputations are formed and whether a reputational signal is credible.

## Reputation Formation

There is little doubt that reputation is important. Much of the theoretical work on reputation stems from repeated play games. Kreps et al. (1982) point to the possibility that agents in a repeated-play prisoners dilemma game can reach one of many cooperating equilibria. Such equilibria require that at least one agent in the population has a reputation as an "irrational" player -- one who does not play the single period Nash equilibrium of defect. Fudenberg and Maskin (1986) generalize the point. Fudenberg and Levine (1992) show that an individual, playing for the long run, might want to invest in a reputation, even if that investment is initially costly (see also Celentani, 1996). They ask whether such an agent would do the same, even if costly actions are only imperfectly observed. Indeed they find that even imperfect observations provide some information about the long-run player's type and from that the long-run player has an incentive to invest in building a reputation, while short-run players have good reasons for drawing inferences about that reputation. As Cripps et al. (1996) show, reputation can dissolve quickly if one or another actor holds incorrect off-equilibrium-path beliefs. However, there is substantial agreement that actors have an incentive to build reputations, regardless of the type of actors they face. Ordinarily these reputations involve some commitment to a form of behavior that is either perfectly or imperfectly observed (although Kim, 1996 points to ways in which “cheap talk” serves to build reputation).

If actors can develop reputations for *trusting and reciprocal behavior*, then in some settings there can be gains to both. How can an initial reputation for trust or reciprocity be constructed? In part it can be accomplished by using observed characteristics of both actors. An initial reputation is often communicated in a variety of ways – through one’s social standing, ethnicity or gender.

It recent years social psychologists have argued that human facial expressions provide critical cues for signaling intentionality. Much of the literature derives from Darwin (1872/ 1998), who argued that humans, like animals, have evolved patterns of signaling behavior, including (but not limited to) facial expressions. The original thrust behind Darwin's characterization is twofold. First, he argues that facial expressions (and other displays) are evolved and serve a function for the species. So a peacock's ostentatious display of his tail or a chimpanzee’s baring of her teeth is innate to the species. Second, expressions serve to signal something to others. In other words, universal, common signals are used to warn, invite or soothe members of the same species and on occasion other species as well.

Contemporary researchers largely follow the lead of Ekman (1972; 1983) and focus on the first part of Darwin's argument. With respect to humans this approach holds that there is a universal set of evolved human facial expressions. Many of these expressions are thought to be involuntary, and reflect basic emotions. The bulk of the research has turned toward understanding what facial expressions reveal about the underlying emotional state of the expressor. As a consequence, facial expressions constitute emotional leakage, in which the emotional content should be obvious to others, since they too share the same universal repertoire of expressions. Of course, learned social behavior works to mask those emotions and cultural differences lead to different forms of masking. Therefore facial expressions can sometime be hard to read (for a general critique of the "universal" recognition of emotion, see Fridlund, 1994, Chapter 10).

In a challenge to what he calls the "emotions view" of faces, Fridlund (1994) takes up Darwin’s second point and proposes a "behavioral ecology" view of faces, arguing that facial expressions and their interpretation by others is crucial. "The balance of signaling and vigilance, countersignaling and countervigilance, produces a signaling 'ecology' that is analogous to the balance of resources and consumers, and predator and prey, that characterize all natural ecosystems." (Fridlund, 1994, p. 128). Facial expressions and their interpretation involve a delicate game in which expressions are signals about intentionality.

*Abstract Images.*

Much of the literature using human facial expressions finds that particular expressions are difficult to "read." That is, the emotional content of an expression is often unclear (see the critique by Russell, 1993). Even something as simple as a "smile" can easily be misinterpreted or misrepresented -- especially if a single snapshot is pulled out of context (Ekman et al., 1998; Leonard et al., 1991; Fernandezdols and Ruizbelda, 1995). To correct for these problems a handful of researchers have adopted highly stylized aspects of faces in order to tease out the primary elements of facial expressions.

If there are specific components of expressions that signal specific emotional or intentional states, then these should be susceptible to systematic evaluation. Taking this insight, McKelvie (1973) designed an experiment in which he used schematic representations of faces. These schematics resemble variations on the ubiquitous "happy face" wishing everyone a nice day. McKelvie used an oval to represent a head and then drew in line segment representations of eyebrows, eyes, nose and mouth. These were systematically varied and then presented as stimuli to subjects.

A total of 128 schematic faces were used; each subject was presented with a sample of 16 faces. Working one at a time, subjects were asked to rate how easy it was to find an adjective to describe the face and then asked to score the appropriateness or inappropriateness of each of 46 adjectives for describing the face. The adjectives reflected four different emotional categories (happy, sad, angry and scheming) and one other category (vacant). His analysis shows that the shape of eyes and the structure of the nose has little effect on evaluations. Instead, eyebrow and mouth shape have the greatest effect. He cautions that neutral (horizontal) eyebrow or mouth expression signals little. "However, when brow and mouth move from the horizontal, clear differences in meaning emerge: medially down-turned brows indicate anger or schemingness; medially upturned brows are seen as sad; an upturned mouth denotes happiness; and a down-turned mouth is seen as angry or sad." (McKelvie, 1973, p. 345). In short, even simple schematic representations of faces can trigger emotional affect that is well recognized.

Part of McKelvie's study was replicated using pre-school children. MacDonald et al. (1996) used schematic drawings of facial expressions thought to represent the six primary emotions. At the same time selected photographs from Ekman and Friesen's "Pictures of Facial Affect" (1978) were used. In one of the experimental conditions children were asked to choose specific emotional categories when viewing either the pictures or the schematics. MacDonald et al. find that accuracy in picking the proper label was significantly greater for the schematic drawings than for the photographs. Accuracy varied, however, with children having the easiest time identifying happiness, sadness and anger (p. 383). The simplifications of the schematics were readily apparent to these children and the emotion evoked was usually readily interpretable. A similar finding for adults comes from Katsikitis (1997) whose subjects compare both pictures of actors and line drawings of those same faces. For certain of the emotions (like surprise), the line drawings tend to be easier to interpret (see also Yamada, 1993).

The lesson to draw from these studies is that humans are very good at recognizing emotional content even in highly stylized schematics. Pictures have meaning and they are readily interpreted. Using these findings we move to two experiments to focus on the behavioral signals that are embedded in simple facial expressions. To avoid the ambiguity associated with still photographs of human faces we use stylized icons to represent facial expressions. Based on the work of McKelvie (1973) and others, we limit our attention to the position of eyebrows and the shape of the mouth.

## Survey and Experiments

Two studies were conducted. The first is a survey designed to reveal subjects' impressions of a series of schematic faces. The survey measures the trustworthiness of the faces as well as the emotional affect attributed to them. The former is a departure from earlier work which has focused almost exclusively on affect; this component of the survey is designed to elicit the influence of facial expressions on subjects' expectations about behavior. We use the survey results to select specific icons for the second study, an experimental test of the effect of schematic faces on the behavior of subjects in simple games involving trust and reciprocity. We are interested in knowing the extent to which the characteristics attributed to the faces shape the behavior of subjects who have been assigned these same faces. In order to test this relationship, we select the two icons from the first experiment that are judged to be the most distinct from each other.

### *The Survey.*

A survey instrument was administered to a sample of 524 subjects (324 male, 192 female and 8 who failed to indicate their sex) in Principles of Economics classes at Virginia Polytechnic Institute and State University in January, 1998. The classes consisted primarily of college sophomores; about 1/3 were business majors, 1/3 engineering majors and the rest from assorted fields. Subjects were asked to complete a three-page survey during a regular class meeting time, either at the beginning or the end of class, and were not compensated for their participation. On the first page of the survey, each subject was assigned one of nine icons and asked to rate its characteristics. The icons are based on a 3x3 design involving three manipulations of the mouth and three manipulations of the eyebrows, and are shown in Figure 3. Only a subset of those icons are analyzed here (for additional discussion see Eckel and Wilson, 1998).

<Figure 3 About Here>

Subjects were randomly assigned to a particular icon and told that the icon “is supposed to represent a type of person.” They then were asked to choose the most appropriate response for their icon on twenty-five word-pair items using a seven-point semantic differential scale. In the scale, a value of (1) means the word on the left is “very” close to matching the meaning of the icon, (2) is “somewhat” close, (3) is “slightly” close and (4) is “neither.” The scale is symmetric to the right of (4). Left/right word order was randomly assigned for the word pairs. In the analysis presented here we focus on ten paired items from the instrument: Five items relate to a behavioral assessment of the icon (does the icon reflect trustworthiness, generosity, cooperativeness, etc.) while the latter five items are common measures of emotional affect (does the icon reflect goodness, happiness, etc.). Figure 4 plots the mean response across four of the icons for each word pair. Each icon's mean for the word pair is connected with a line running down the graph. The icons at the top of the graph match the order of the means for the trusting/suspicious word pair. From left to right, the *happy* face is judged on average to be more trusting and trustworthy than the *sad* face, the *angry* face and the *devious* face. What is clear from the figure is that the order is preserved across nearly all ten items.

<Figure 4 about here>

“Eyeballing” the means pretty much tells the story. Across the first five “behavioral” items, the differences among icons are significant under pair-wise t-tests (the only exception is between the “angry” and “devious” icons on the generous/selfish item). In other words, subjects perceive the icons as representing different behavioral traits. The same is true with respect to affective (emotive) items. A similar ordering is preserved, with the "happy" icon being the most positively evaluated. The "sad" icon is generally regarded positively, except on the happy/sad item, where it is appropriately rated by subjects. Finally, the "angry" and "devious" icons switch positions on the good/bad, kind/cruel and friendly/unfriendly items. However, this is consistent with what is ordinarily thought of as angry and devious affective states. Moreover, these findings are consistent with those found by other researchers who have used either schematic or facial images.

The first five items and last five items were combined to generate two scales. Both the “behavior” and the “affect” scale add an individual respondent’s score across the items contained in Figure 4. An average score across these items was then calculated for each respondent. Two separate models were then estimated for each scale as shown in Table 1. Four independent variables are included as well as an additional control variable; Model 2 includes interaction terms. The variable SMILE is a dummy variable for icons with an upturned mouth. Likewise FROWN is a dummy variable for a down-turned mouth. UPBROW is a dummy variable for eyebrows that are upturned at the center (/ \)and DOWNBROW does the same for down-turned or "frowning" eyebrows (\ /). The neutral mouth and brow are then reflected in the intercept term of the regression. In model 2, these variables are interacted as indicated. Finally we add a dummy variable for the SEX of the respondent, controlling for perceptual differences that might emerge from assessing these icons.

<Table 1 About Here>

Both models confirm findings by McKelvie (1973) and others. The positioning of both the mouth and eyebrows significantly affects the assessment of the facial icons. With respect to the behavioral scale (trust, honesty, etc.) the intercept term reflects the midpoint of the general semantic differential scale and is consistent with what we might expect from a neutral icon. The effect of eyebrows is pronounced. When the eyebrows are upturned, they decrease the evaluation (move it toward the “trustworthy” end of the scale) by almost a full point. Down-turned eyebrows have the opposite effect. Interestingly, the mouth position does not affect behavioral assessments except in interaction with the eyebrow positions. In Model 2, the interaction between Smile and Downbrow has a strong positive effect on the index, indicating a move toward the “untrustworthy” end of the scale. A smile does not have an unambiguous effect on the perception of the likely behavior of a facial icon. The effect can be positive or negative, depending on the position of the eyebrows. Main effects alone would lead one to believe that the frown/downbrow combination would be most negatively perceived, but interaction effects adjust the evaluations so that the conflicting message of the smile/downbrow icon is perceived as suspicious and dishonest.

When turning to the affect scale, both the smile and the eyebrow positions are strongly related to the evaluation of the icon. These results are consistent with the behavior scale in that a smile and upturned eyebrows result in a more positive assessment, while a frown and down-turned eyebrows lead to a more negative assessment. Interaction terms are again strong, with the combination of a smile and downturned brows having a significant effect on the evaluation of the icon. What also appears from the estimation is that female respondents are more likely to evaluate the icons harshly with respect to affect. The effect is not large, but is statistically strong.

In short, these estimations reinforce the results displayed in Figure 4. The position of eyebrows and mouth both matter for the inferences that respondents draw about the icon. Put another way, respondents are drawing meaning from the icons, and the meaning is systematically related to our manipulations. Subjects differentiate between the various icons, and that the icons can be ordered from most to least *positive* on both behavioral and affective characteristics. These results set the foundation for subsequent experiments.

*Experiments*

In the experiments reported here pairs of subjects participate in series of two-person sequential games. All games involve sequential choice under perfect and complete information. In a typical game each subject makes one or two moves. A total of 168 subjects were recruited from the local student population at Rice University. Students were contacted in their dining halls and asked to volunteer for a decision making experiment. Subjects signed up for one of sixteen planned experimental sessions. (Fourteen sessions were conducted in February, 1998; an additional nine were added in November, 1998).

The laboratory accommodates eight subjects, each seated in a cubicle formed by moveable partitions and facing a computer. Although subjects are in the same room and can hear one another, they cannot see one another's computer screen. At the outset of the experiment subjects are cautioned not to speak and told that if they do so, then the experiment will be canceled. All experimental sessions are conducted over local area network and all communication between subjects is handled by the network.

Upon arriving at the laboratory subjects choose their seats and are asked to sit quietly until all the volunteers have arrived. At least nine subjects are recruited for each session in order to ensure that eight participate. If more show up, a volunteer is solicited and paid $3.00 on the spot to leave. If no one volunteers, one subject is randomly selected, paid and dismissed. Once the requisite number of subjects appears, the experiment begins. In eight of the twenty three sessions fewer than eight turned up, and only six subjects participated in the experiment (only an even number of subjects could be used in this experiment).

Subjects are given self-paced instructions and shown how to indicate their choices in the sequential game.[2] These instructions are attached as Appendix 1.[3] In each

[2] The task for the subjects was referred to as a "decision problem." The term "game" was avoided, but is used throughout the text to denote a decision problem characterized as a game theoretic problem.

experimental session, subjects participate in as many as 30 games.[4] Prior to each decision subjects are randomly rematched. Because of the limited number of subjects, same-pair play often occurs. However, subjects carry no unique identification in the course of the experiment, so it is impossible for subjects to know with whom they are paired in each game.

*Games.* A set of 18 distinct games is used during the course of the experiment. For the first six periods of play the games are presented in a fixed order. In all subsequent periods games are randomly selected for each pair in each period. Therefore in our analysis of this design we have to be careful to account for history effects that are a function of the order of play.

Only seven of the 18 games are discussed here. These games are given in Figure 5 and were selected to tackle two concerns. Games 1, 2 and 3 were chosen to assess whether subjects understood basic properties of backward induction in this experimental design. Aside from having a single branch, games 1 and 2 share the property that the equilibrium is obvious. In these games subjects have no incentive to use strategies that yield out-of-equilibrium behavior. Neither trust considerations nor concerns about reciprocated trust enter into a subject’s strategic calculation when playing these games. Game 3 is a two-branch game and is a bit more complex. Unlike Games 1 and 2, it requires more calculation on the part of subjects, has two subgames and contains two Nash equilibria which are at the final node for each branch. Overall the unique subgame perfect equilibrium is the last node on the left branch.[5] Games 1, 2 and 3 share the characteristic that there is no incentive for subjects to do anything except to play their narrow self-interest.

<Figure 5 About Here>

Games 4, 5, 6 and 7 were selected to test for concerns with trust and reciprocity. In each game the Nash equilibrium involves the first player quitting at the first node. However, it is clear that if the first player can read the intentions of the second player, trust that player by passing, and then have that trust reciprocated, the first player is left better off. At the same time, the second player, by evoking a trusting move, is left better off then

[3] On average subjects took 7 minutes to complete the instructions, with no subject taking longer than 13 minutes. In post-experiment debriefings subjects remarked that the instructions were clear and said they had no problem with the task.

[4] In the first experimental session, because we were uncertain about the length of time required to complete the games, subjects participated in only 20 periods. In the second session, this was raised to 25 periods. In both sessions subjects were debriefed and asked whether the task was too boring or repetitious. Subjects indicated they were not bored and would not have minded participating in more decisions. All subsequent periods used 30 periods. Each game usually lasted less than a minute.

[5] To ensure there was not a selection bias, branches to this game were randomly flipped throughout the experiment.

if receiving the Nash equilibrium. However, by reciprocating that trust, the second player gives up a temptation payoff. For game 6 this can mean as much as $30. A useful way of thinking about each of these games is to compare across games the values to player 1 of trusting (*gain),* and to player 2 for defecting (*temptation).* For game 6, *gain* is the difference between Player 1's maximum payoff at the final node and his payoff at the first node, or 20-9=11. *Temptation* is the difference between Player 2's payoff at node 2, and at the final node (assuming Player 1 maximizes his payoff at the final node); for game 6, this is 30-10=20. Games 4-7 differ in these measures, illustrating both the benefit to trust and the different costs of reciprocating trust.

*Procedure*. The same procedure is used in each period of play. Before each game begins, subjects are told which roles they are assigned (1 or 2) and who will move first. Each subject is randomly assigned to be player 1 or 2 in each period (in the experiment, we refer to "Decision Maker" 1 or 2). Information about a player's counterpart is then revealed, depending on the icon manipulation as explained below. In the games we analyze here, player 1 always has the first move and player 2 waits until the first player's choice is complete. At each of the first two nodes a player has two choices -- either to quit the game and take the payoffs or to pass to the next player. Once a choice is made, the other player is notified of the move. If the choice is to quit, the payoff box is circled on the computer screen, and both players are notified of the outcome and asked to record their payoffs for that period. If the move is passed, the second player must make a choice while the first player waits. All communication between players is mediated by the computer and subjects are only told the moves of their own partners. If the game continues to the last decision node, the first player has the choice of two payoff boxes. Once the game ends, subjects are instructed to wait until all pairs have completed their own decisions. At that point the subjects are again shuffled and re-paired.

Participants were paid in cash and in private at the conclusion of the experiment. All payoffs in the games in Figure 5 are in U.S. dollars. Subjects are told at the outset that they would be paid only for a single period of play. At the conclusion of the experiment they are asked to draw one card from a deck of 100 electronic cards displayed on the computer screen. Subjects are told that each period has an equal probability of being chosen, and the algorithm for the selection ensures this. No matter which card is turned over, the program randomly selects a period and informs the subject of the period drawn and earnings. Subjects are asked to verify that payoffs for the period drawn match what they record. Before they are paid, subjects fill out an on-line questionnaire that asks them questions about their participation. No experimental session lasted more than 60 minutes and none was shorter than 40 minutes. On average subjects earned $13.47 for their participation. One subject earned the maximum of $30.00 and seven subjects earned $0.00 for their play. These latter subjects were paid a show-up fee of $3.00, but not informed until their debriefing that they would be paid this amount.

*Icon manipulations.* Subjects know little about their partners in each trial. At best they know they are paired with other subjects in the same room, but they never know the identities of their partners. In order to build a reputational signal, however, each subject is assigned a permanent identity at the outset of the experiment. Five distinct manipulations of identity are used. In each session half of the subjects are randomly assigned one icon representation and the remaining half are assigned the other. Figure 6 presents the icon pairs used in the experiment. Subjects know there is a population of two types of icons in experiment. They know that each subject had been assigned an icon at the outset and that each player retains his or her assigned icon for the duration of the session.

We are very deliberate in not tying the icon to any personal characteristics of subjects. They are simply told the icon assigned to them will be theirs for the entire experimental session. Our sense is that this constitutes a very weak stimulus, and that a stronger connection between the icon and the subject would strengthen any observed behavioral effects.

<Figure 6 About Here>

The primary manipulations for the experiment involve the pairings of icons with and without human facial characteristics. Sessions are of three types: those with facial icons, with nonfacial icons, and with no icons. As noted above and detailed below, we have explicit predictions about non-game-theoretic play in each of the games. These behaviors are influenced by the icon type of a player and the partner. Two facial icons were used in which the angle of the eyebrows and orientation of the mouth were changed. The icon with downturned eyebrows and an upturned mouth is characterized as “devious;” the icon with upturned eyebrows and an upturned mouth is characterized as “happy.” Two additional icons are used that have no human facial content. Here a rectangle and an oval are paired. These icons are chosen as one control condition for the experiment. It may be that subjects do not use the human facial content in the icons to select strategic play. Instead they simply view the world as consisting of two types – “them” and “us”. “In-group/out-group” effects are common in social psychological experiments (Tajfel and Turner, 1979; Turner, 1978), and we might reasonably predict that subjects identify with their own icon type and play differently then when confronting a different icon type.

In each trial a subject can be paired either with an individual with the same icon or an individual with a different icon. Prior to beginning a game subjects are shown the entire set of icons in the game (for an 8-person group this meant four images of each type of icon). When the subject is ready to begin, the icons appear to be shuffled on the screen and one is selected by the program. The screen then displays the subject's own icon and the icon of his or her counterpart. When the subject chooses to continue, the game is then displayed.

A final control condition was also introduced in which subjects have no information about their counterparts. Their "icons," for all intents and purposes, are blank (9). They are not presented any screens in which they are told about their counterpart's identity, but simply play a series of games in which they are told they have been randomly matched with another participant. This control group is designed such that no reputational content is provided. Subjects in this condition should exhibit behavior that more closely approximates the predictions of standard models of game theory.

*Predictions*

Given the earlier discussions and the detailed discussion of the experimental manipulations, several distinct predictions are offered.

> *Prediction 1*: Subjects will choose the subgame perfect equilibrium, irrespective of the game.

The standard game theoretic prediction is that subjects will use subgame perfection. In Figure 5 above, the subgame perfect equilibrium is circled for each game. Despite the fact that subjects in some conditions are given an identity, the standard game theoretic model treats these icons as uninformative. In addition, because these icons are randomly assigned to subjects and subjects are given no specific rationale for that assignment, there should be no reputational content. On the other hand, our discussion above leads us to predict that in games 1, 2 and 3 subgame perfection will predict the outcome. In those games there is nothing to gain from reading the intentions of others. By contrast, in games 4, 5, 6 and 7 we expect subjects to engage in trusting and reciprocated trustworthy behavior. This will lead to consistent out-of-equilibrium outcomes. This gives rise to our second prediction.

> *Prediction 2*: In games 4, 5, 6 and 7 trust and reciprocity will vary across manipulations.

In all of these games a trust move is a move by player 1 to forego the subgame perfect equilibrium (which is to immediately quit) and pass the move to player 2. A reciprocal trust move is then made by player 2 to pass the move back to player 1 rather than choosing to quit. The degree of trusting and trustworthy behavior, however, will vary with settings that generate reputational content and those that do not. Therefore, we predict much greater trust and reciprocity in manipulations with facial icons than in manipulations without facial icons. Experimental conditions in which subjects are given rectangles and ovals are expected to convey no reputational information. A good deal of literature in social psychology points out that in-group and out-group effects are easily generated and so it is plausible that when any pair of subjects have the same icon, they will engage in trust and reciprocity. This constitutes a rival prediction.

We also expect each of the games to have an independent effect on trust and reciprocity. The discussion above noted that actors choose strategies contingent on a number of factors, including the reputation of their counterparts as well as the context of the game. Games 4, 5, 6 and 7 all differ with respect to player 1's *gain* player 2's *temptation*. If player 2 passes the move back to player 1 and that individual chooses his most preferred outcome, then it is simple to calculate the value of the temptation move compared with to the reciprocated move as described above. Consequently, the context of the game itself ought to have an effect on the extent of trust and reciprocity, with a decline in such behavior as the size of the temptation relative to gain increases.

*Prediction 3*: Initial reputation matters for trust.

The facial icons assigned to subjects are assumed to carry reputational content. Drawing on the findings from the first experiment, subjects are expected to draw specific inferences about each of the icons. In games 4, 5, 6 and 7, player 1 will be sensitive to player 2's icon in deciding whether to make a trust move. The results from the first experiment clearly show that the "happy" icon is most likely to be trusted, and "devious" the least likely to be trusted. Therefore, we expect systematic differences in the decisions by player 1 given their counterpart's icon.

*Prediction 4*: Reputation formation matters.

Given the design of this experiment, in which subjects participate in many different games with a limited population of icon "types," it is possible to disentangle whether subjects develop reputations. These icons are predicted to have an initial set of characteristics that are commonly understood by subjects. At the same time, given the limited population of icons, it is possible that they take on different meanings over the course of the experiment. Our hypotheses is that facial icons facilitate reputation-formation. Subjects interpret the play of others with the "happy" icon differently than the play of others with the "devious" icon. Thus "happy" icons are more likely to develop a positive reputation and "devious" more likely to develop a negative reputation. While rectangle and oval icons, which have no facial content, may become signals of reputation, there is no reason for one or the other to develop positively or negatively.

*Analysis.*

First we ask whether subjects understand the basic structure behind the game. As noted above, games 1, 2 and 3 are very straightforward. There is no reason in these games to assess the intentions of the other and there are no reasons for engaging in trust. Consequently the subgame perfect equilibrium for the game ought to be the predicted outcome. Figure 7 provides the descriptive statistics for the node at which subjects ended the game. Consistent with Prediction 1, subjects consistently choose equilibrium play across games 1 and 2. In game 1 it is chosen 94 percent of the time and 91 percent of the time in game 2. By almost any standard, this points to the fact that subjects understood the underlying structure of the experiment and behaved consistent with game theoretic expectations. Under game 3 the obvious first choice was for player 1 to go left -- thereby obtaining the largest payoff. Both players should then continue to pass until reaching the bottom node. In this more complex game (requiring that two branches be compared, subjects chose the equilibrium 81.3 percent of the time. Another 12.5 percent of the time player 1 chose the right branch and in that instance all subjects went to the equilibrium for that subgame. Only a handful (3 subjects) made anything equivalent to a mistake -- that is, an outcome inconsistent with equilibrium play. All-in-all, subjects appear to understand backward induction and behave according to game theoretic predictions.

<Figure 7 About Here>

By contrast, subjects in games 4, 5, 6 and 7 exhibit far less equilibrium behavior and considerable trusting and trustworthy behavior. Figure 8 contains the aggregate results for these games. In Game 4, for example, Player 1 moves right 38.8 percent of the time, and trusts 61.2 percent of the time. About half of the trust moves are reciprocated, with 32.9 percent of the subject pairs reaching the third node.

It is apparent from the data in Figure 8 that Prediction 1 can easily be rejected for Games 4-7. Almost two-thirds of subjects deviate from game-theoretic play: pooling all games, 62.7 percent trust at node 1 of the games, and 51.9 of those who were trusted reciprocate at node 2. At the final node, subjects overwhelmingly choose the “equal-split” option, rewarding Player 2’s reciprocity in 85.6 percent of the decisions that reached the third node of the game. On the other hand, subjects do not uniformly trust or reciprocate, though the majority of them do. Clearly the overall results indicate that subjects sometimes do, and sometimes don’t, trust and reciprocate. Further analysis gives some insight into when greater trust and reciprocity might be expected.

A comparison of the games gives some insight into the effect of the differences in the payoff structures of the games. Compared with Game 4, Game 5 has a larger *gain* with a similar *temptation,* and there we see a higher level of trusting behavior (71.1 percent of first moves are trusting.) Games 5 and 6 have similar levels of *gain*, but Game 6 has a greater *temptation*. Here we see an increase in defections, with about 2/3 of the trust moves ending in the defection of Player 2. In addition, Player 1 is able to anticipate the defections, and trusts less frequently. Game 7 has the lowest *gain*, and we observe the least amount of trust in this game. Both *gain* and *temptation* appear to affect subjects' play, consistent with Prediction 2.

<Figure 8 About Here>

To elaborate on Prediction 2 we provide Table 2. The first column contains the icon of the first-mover; the first row shows the icon of the second-mover. Numbers in the table show the proportion of trusting moves by pairing. This table shows that the degree of trust is higher when facial icons are present than when nonfacial icons are present. Sessions with no icons on average show an intermediate level of trust. In addition, a small in-group effect appears in the data, with both facial and nonfacial icons more likely to trust icons of their own type.

Table 3 contains probit regressions analyzing the probability that Player 1 will take a trusting move. The variable FACE is equal to 1 if a facial icon is present. BLANK takes on a value of 1 for sessions with no icons. SPRING takes on a value of 1 for the sessions that took place in February, 1998. BLANK and SPRING are also interacted. Model 1 shows that, in sessions where icons with facial characteristics are present, subjects are more likely to take an initial trusting move than in the sessions with non-facial icons. BLANK sessions exhibit a high degree of heterogeneity, with SPRING/BLANK sessions significantly less trusting than in the fall sessions. Because of the inconsistency in the Blank data, they are dropped from subsequent analysis. Model 2 contains data from sessions with icons only. Note that the coefficient on FACE is large and statistically significant, indicating a greater degree of trust in those sessions. The coefficient on SPRING is also positive and significant, perhaps indicating a subject pool difference. Model 3 tests for ingroup effects. The four new variables in this model take on values of 1 when the relevant icon is paired with one of the same type. For example, PairDevious is 1 when Devious icons are paired with other Devious icons. The pattern of coefficients would seem to indicate a significant in-group effect only for facial icons. Finally, model 4 re-estimates model 2, including the measure of the first mover's anticipated gains from trusting. This coefficient is positive and significant, pointing out that subjects are sensitive to price.

<Table 3 About Here>

Results in Table 3 further support Prediction 2: there appear to be significant differences in behavior across treatments, with stronger trust appearing in the sessions with facial icons.

The first play of Game 4 is worth special attention, because this is the very first game that subjects play, and so is not affected by previous history. Prediction 3 suggests that players will use facial icons to form initial priors, and that will affect initial play. In Table 4 we report probit regressions for the very first play of the experiment (55 observations). The dependent variable in the Trust model is the probability that Player 1 takes a trusting move. In the Trust equation, Devious is equal to 1 when Player 2 (the partner of Player 1) displays a devious icon; Happy is one when Player 1 faces a happy icon; Rectangle is 1 when Player 1 faces a Rectangle icon. None of the variables approaches statistical significance in the Trust equation. The dependent variable in the Reciprocity equation is the probability that Player 2 moves down at the second node, reciprocating Player 1's trusting move. The variables now represent the icons that Player 2 faces. Again none of the variables are significant, although Happy approaches significance. From these equations there is no clear evidence that the icons affect initial play; Prediction 3 is not supported by the data.

<Table 4 About Here>

Prediction 4 suggests that the development of reputations will be affected by the play of the game and by the icons that are present. These patterns are analyzed in more detail in Tables 5 and 6, where we incorporate game history into the analysis. In Table 5 we analyze trust, and in Table 6, reciprocity. The purpose of this analysis is to examine the effect of the facial and nonfacial icons on reputation formation, controlling for the history of the game that a player experiences. For this analysis we again pool all decisions over all plays of games 4-7.

In Table 5, the dependent variable takes on a value of 1 if Player 1 trusts by moving down at the first node of the game. Devious, Happy, and Rectangle represent the icons that player 1 faces; if Player 1 faces a Devious icon, Devious takes on a value of 1. GAIN is the measure introduced above of the potential gain to trust and reciprocity; it is the difference between Player 1's payoff if he chooses not to trust by moving right at the first node, and his maximum payoff at the third node of the game. Model 1 indicates that both Happy and Devious icons elicit more trusting behavior than Rectangle or Oval (the intercept). GAIN also significantly affects trust, with greater potential gain inducing more trusting behavior.

<Table 5 About Here>

In Model 2 we add three additional variables that reflect the history of the game. BETRAYED measures the percentage of the time when the player has been betrayed by a partner. This measure is developed over all games in the experiment, not just those analyzed here, and so is a summary measure of the fraction of time that this player's trust has been violated. If the player's trust was always betrayed, this variable would take on a value of 100; if never, zero. Our prediction is that the more a player's trust is betrayed, he is less likely to trust in the future. The variable BETRAYER measures the fraction of time that the player has betrayed a partner. Again, this measure is developed for the entire history of the game. EARNINGS captures the subjects' earnings up to the point of the current decision, and is a measure of "success" in playing the games. Model 2 shows the same pattern of results as in Model 1 for the variables they have in common. BETRAYED is insignificant and carries the opposite sign from our expectations. This may be because this variable is also picking up the subject's propensity to trust, which might be resistant to experience. (We are working on refining this measure.) BETRAYER on the other hand, is highly significant. This indicates that people who betray the trust of others are not likely to be trusting of others. Earnings carries the expected positive sign, but is not significant.

Table 6 contains similar estimates for the decision to reciprocate. The dependent variable is 1 if Player 2 reciprocated a trusting move by moving down at the second node of the game. Model 1 is similar to Model 1 in Table 5, with the icon variables representing the icon of the player who is paired with Player 2. Temptation is the difference between Player 2's payoff at node 2, and her lowest payoff at node 3. In Model 1, Temptation is strongly significant, and there is a pattern of significance in the icon variables. In Model 2 the additional history variables are included. Betrayers are more likely to betray a trusting move, and those with higher earnings are more likely to reciprocate. Icons lose statistical power. Finally, in Model 3 we add a variable that represents Player 2's experience in the previous game. LAST EXPERIENCE is 1 if Player 2 was betrayed in the immediately preceding game. This experience has a powerful influence on Player 2's decision.

<Table 6 About Here>

*Discussion.* We have presented analysis of experimental data on the effect of facial icons on trust and reciprocity. Our primary finding is that in the presence of randomly-allocated facial icons there is both more trust and somewhat more reciprocity (though the evidence on reciprocity is weak) than when subjects are presented with nonfacial icons. This effect emerges as reputation unfolds. We do not find a significant difference in the behavior of the two facial icons, perhaps because they are always paired in a session. In similar research using the ultimatum game, we do find significant differences in behavior depending on the subjects' own icon and the icon of his partner. Additional research is necessary to disentangle the effect of facial icons per se from that of particular facial expression.

## Conclusion

Individuals are neither always altruistic nor always coldly calculating of their narrow advantage. Instead individuals respond to their environment in a contingent manner, assessing not only the potential gains from their own strategic behavior, but also assessing the intentions of others and the strategies they will play. While incorporating the strategic play of other players lies at the heart of game theory, as a behavioral description it does not fit observed interactions in many contexts. Mathematical tractability has led theorists to generate models that more or less describe autistic behavior -- the mental state of other players is irrelevant. An actor need only turn inward and ask how she would play if she was in the other's position.

These experiments have not established the foundation and source of reputation. However, what is clear from these data is that subjects do use reputational cues. Those cues come directly from simple elements of facial characteristics. When facial characteristics are absent, subjects exhibit less willingness to "trust" their partner. When those characteristics are present, then trusting behavior is more evident.

The fact that simple facial cues affect behavior poses a challenge for game theory. On the one hand, game theoretic predictions do quite well in the absence of reputational priors. This is encouraging because it tells us that game theoretic models appear to be on the right track. On the other hand, those same predictions do miserably when reputational priors are randomly generated. What is especially worthy of note is that only icons with facial characteristics generate reputations. This implies that game theory, which nominally is about the strategic interaction of human beings, needs to pay attention to human beings. As it now stands, for some forms of two-person interaction, standard game theoretic models appear to be a model of autism -- strategic interaction that only looks inward and fails to acknowledge the presence and active participation of others.

These experiments buttress the findings of many others. Simple facial schematics embody affective content. Our results also point out that these schematics embody social content as well. They are used, at the margin, to generate trust and reciprocity. This finding moves beyond a simple accounting of emotional content to point out that the inferences drawn by subjects have behavioral  characteristics.

Generally these results show that simple cues can trigger reputational priors and those priors have real content.  They result in more or less cooperation in settings in which cooperation is costly.  Whether the "trigger" is due to standard concepts of stereotyping, attribution or evolutionary psychology remains an open question.  Nonetheless these findings are intriguing and call for additional work.

## Bibliography

Aronoff, Joel, Andrew M. Barclay and Linda A. Stevenson. 1988. “The Recognition of Threatening Facial Stimuli” *Journal of Personality and Social Psychology*. 54: 647-655.

Aronoff, Joel, Barbara A. Woike and Lester M. Hyman. 1992. "Which Are the Stimuli in Facial Displays of Anger and Happiness? Configurational Bases of Emotion Recognition." *Journal of Personality and Social Psychology*. 62: 1050-1066.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Baron-Cohen, Simon. 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* Boston: MIT Press.

Berg, Joyce; Dickhaut, John W.; McCabe, Kevin A. 1995. “Trust, Reciprocity, and Social History.” *Games and Economic Behavior*; v10 n1, July, pp. 122-42.

Budesheim, T. L. and S. J. DePaola. 1994. "Beauty or the beast? The effects of appearance, personality and issue information on evaluations of political candidates." *Personality and Social Psychology Bulletin*. 20: 339-348.

Camerer, Colin, 1998. “Simple Bargaining and Social Utility: Dictator, Ultimatum and Trust Games.” Chapter 3 in *Experiments in Strategic Interaction*, forthcoming.

Cherulnik, P. D. , L. C. Turns and S. K. Wilderman. 1990. "Physical appearance and leadership: Exploring the role of appearance-based attribution in leadership emergence." *Journal of Applied Social Psychology*. 20: 1530-1539.

Cosmides, Leda and JohnTooby. 1992. “Cognitive Adaptations for Social Exchange.” In Barkow, Jerome H., Cosmides and Tooby, eds., *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* 1992. New York: Oxford University Press.

Eckel, Catherine C. and Philip Grossman, “Altruism in Anonymous Dictator Games.” *Games and Economic Behavior 16*:181-191, 1996.

Eckel, Catherine C. and Philip Grossman, “Are Women Less Selfish Than Men?: Evidence from Dictator Games.” Forthcoming, *The Economic Journal*, May, 1998.

Darwin (1998) -- *The Expression of the Emotions in Man and Animals* (Series in Affective Science) by Charles Darwin. Paul Ekman (Editor) Oxford University Press.

Ekman, Paul (1972) -- *Emotion in the Human Face : Guide-Lines for Research and an Integration of Findings*. (Pergamon general psychology series ; PGPS-11) New York: Pergamon Press.

Ekman, Paul. 1982. *Emotion in the human face*, 2nd ed. (Studies in emotion and social interaction) New York: Cambridge Unviersity Press.

Ekman, Paul. 1997. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (Series in affective Science) New York : Oxford University Press.

Ekman, P. and W. V. Friesen. 1978. *The Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., W. V. Friesen and M. O'Sullivan. 1998. "Smiles when lying." *Journal of Personality and Social Psychology*. 54: 414-420.

Fernandezdols, J. M. and M. A. Ruizbelda. 1995. "Are smiles a sign of happiness -- Gold Medal winners at the Olympic games." *Journal of Personality and Social Psychology*. 69: 1113-1119.

Fehr, Ernst, Georg Kirchsteiger, and Arno Reidl. 1993. “Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics* 108(2): 437-59.

Forsythe, R., Horowitz, J. L., Savin, N. E. and Sefton M. (1994). ‘Fairness in simple bargaining experiments.’ *Games and Economic Behavior*, vol. 6, pp. 347-369.

Frank, Robert. 1997. “Nonverbal Communication and the Emergence of Moral Sentiments.” In Ullica Segerstrale and Peter Molnar (eds.) *Nonverbal Communication: Where Nature Meets Culture*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp. 275-292.

Fridlund, Alan J. 1994. *Human Facial Expression: An Evolutionary View*. San Diego: Academic Press.

Fundenberg, D. and E. Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica*. 54: 533-544.

Fudenberg, Drew and David K. Levine. 1992. “Maintaining a Reputation When Strategies Are Imperfectly Observed.” Review of Economic Studies. 59 (3): 561-579.

Gopnik, A. 1993. Mindblindness. Unpublished essay, University of California, Berkeley. Cited in Baron-Cohen (1995).

Hamilton, W. D. 1964. “The Genetical Evolution of Social Behaviour (I and II).” *Journal of Theoretical Biology* 7: 1-16, 17-52.

Hoffman, E., McCabe, K., Shachat, K. and Smith, V. (1994). ‘Preference, property rights and anonymity in bargaining games.’ *Games and Economic Behavior* , vol. 7, pp. 346-80.

Hoffman, E., McCabe, K. and Smith, V. (1996). ‘Social distance and other-regarding behavior in dictator games.’ *American Economic Review*, vol. 86, pp. 653-60.

Hoffman, E., McCabe, K. and Smith, V. (forthcoming) “Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology.” *Economic Inquiry*.

Johnson, Mark H., Suzanne Dziurawiec, Haydn Ellis and John Morton. 1991. "Newborns preferential tracking of face-like stimuli and its subsequent decline." *Cognition*. 40: 1-19.

Katsikitis, Mary. 1997. "The classification of facial expressions of emotion: a multidimensional scaling approach." *Perception*. 26: 613-626.

Kreps, D.M., J.D. Roberts, P. Milgrom, and R. Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *Journal of Economic Theory*. 27: 245-252.

Le Doux, Joseph 1996 *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* New York: Simon and Schuster.

Leonard, C. M., K. K. S. Voeller and J. M. Kuldau. 1991. "When's a smile a smile? Or how to detect a message by digitizing the signal." *Psychological Science*. 2: 166-172.

MacDonald, P.M, S.W. Kirkpatrick and L. A. Sullivan. 1996. "Schematic Drawings of Facial Expressions for Emotion Recognition and Interpretation by Preschool-Aged Children." *Genetic, Social and General Psychology Monographs*. 122: 375-388.

Masters, R. D. and D. G. Sullivan. 1989. "Facial Displays and Political Leadership in France." *Behavioral Processes*. 19: 1-30.

McKelvie, Stuart J. 1973. “The meaningfulness and meaning of schematic faces.” *Perception and Psychophysics* 14 (2): 343-348.

MacRae, C. Neil, Charles Stangor, and Miles Hewstone (Editors), *Stereotypes and Stereotyping*, Guilford Press, 1996

O’Connell, Sanjida. 1998. *Mindreading: An Investigation into How We Learn to Love and Lie.* New York: Doubleday.

Ostrom, Elinor. 1998. “A Behavioral Approach to the Rational Choice Theory of Collective Action.” American Political Science Review. 92: 1-22.

Pinker, Steven. 1994. *The Language Instinct.* New York: Harper-Collins.

Pinker, Steven. 1997. *How the Mind Works*. New York: W. W. Norton & Co.

Remnick, David. 1998. *King of the World : Muhammad Ali and the Rise of an American Hero.* New York: Random House.

Russell, J. A. 1993. "Forced-Choice Response Format in the Study of Facial Expression." *Motivation and Emotion*. 17: 41-51.

Sorce, J. F., R. N. Emde, J. J. Campos and M. D. Klennert. 1985. "Maternal emotional signaling: Its effects on the visual cliff behavior of 1-year-olds." *Developmental Psychology*. 21: 195-200.

Sullivan, Denis G. and Roger D. Masters. 1988. "'Happy Warriors': Leaders' Facial Displays, Viewers' Emotions, and Political Support." American Journal of Political Science. 32: 345-368.

Tajfel, Henri and J. Turner, "An Integrative theory of intergroup conflict," pp. 33-47 in W. G. Austin and S. Worchel (eds.) *The Social Psychology of Intergroup Relations*. Monterey, CA: Brooks/Cole (1979).

Tooby, John and Leda Cosmides. 1992. "The Psychological Foundations of Culture." In Jerome H. Barkow, Leda Cosmides and John Tooby (eds.). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Cambridge: Oxford University Press, pp. 19-136.

Trivers, R. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35-57.

Turner, John, "Social Categorization and Social Discrimination in the Minimal Group Paradigm." in Tajfel, Henri. *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. London: Academic Press Inc. (1978)

Wright, Robert. 1994. *The Moral Animal: Why We Are the Way We Are: The new Science of Evolutionary Psychology*. New York: Vintage Books.

Yamada, Hiroshi. 1993. "Visual Information for Categorizing Facial Expression of Emotion." *Applied Cognitive Psychology* 7: 257-270.

Zebrowitz, Leslie A. 1997. *Reading Faces: Window to the Soul?* Boulder, CO.: Westview Press.

## Table 1
## Effect of Facial Characteristics on Perceptions of Icons
## Standard Errors in Parentheses

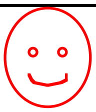
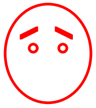
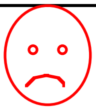
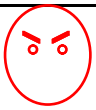
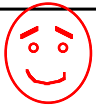
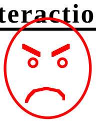
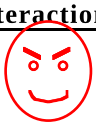
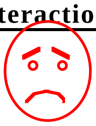
| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | Behavior | Affect | Behavior | Affect |
| Intercept | 4.11**<br>(.10) | 4.10**<br>(.10) | 4.15**<br>(.14) | 4.05**<br>(.14) |
| <br>SMILE | -.09<br>(.10) | -.78**<br>(.10) | -.34°<br>(.19) | -.96**<br>(.18) |
| <br>UPBROW | -.80**<br>(.11) | -.57**<br>(.11) | -.67**<br>(.18) | -.42*<br>(.18) |
| <br>FROWN | .13<br>(.10) | .70**<br>(.10) | .27<br>(.19) | 1.04**<br>(.18) |
| <br>DOWNBROW | 1.02**<br>(.10) | 1.14**<br>(.10) | .87**<br>(.16) | 1.17**<br>(.16) |
| SEX<br>(1=Female) | .05<br>(.09) | .19*<br>(.08) | .04<br>(.08) | .18*<br>(.08) |
| <br>Interaction | | | -.21<br>(.26) | -.09<br>(.26) |
| <br>Interaction | | | -.33<br>(.24) | -.69**<br>(.24) |
| <br>Interaction | | | .88**<br>(.24) | .62**<br>(.24) |
| <br>Interaction | | | -.17<br>(.26) | -.35<br>(.26) |
| r2 | .40 | .51 | .45 | .54 |
| | **p<.01 | *p<.05 | °p < .10 | |

**Table 2: Percentage of First-Movers Choosing to Trust**
**(All games, all series pooled; numbers in italics)**

| First Mover's Icon | Second Mover's Icon | | | | | |
|---|---|---|---|---|---|---|
| | 😈 | 🙂 | □ | ○ | **Blank** | **Total** |
| 😈 | 72.9<br>*62/85* | 54.8<br>*46/84* | - | - | - | 63.9<br>*108/169* |
| 🙂 | 65.7<br>*69/105* | 69.3<br>*61/88* | - | - | - | 67.4<br>*130/193* |
| □ | - | - | 50.0<br>*20/40* | 38.4<br>*15/39* | - | 44.3<br>*35/79* |
| ○ | - | - | 45.7<br>*43/94* | 60.5<br>*26/43* | - | 50.4<br>*69/137* |
| **Blank** | - | - | - | - | 69.1<br>*103/149* | 69.1<br>*103/149* |
| **Total** | 68.9<br>*131/190* | 62.2<br>*107/172* | 47.0<br>*63/134* | 50.0<br>*41/82* | 69.1<br>*103/149* | 62.7<br>*445/727* |

## Table 3
## Probit Estimates for "Trusting" in Games 4-7
## (Standard Errors in Parentheses)

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| INTERCEPT | -0.046<br>(.085) | -0.137<br>(.090) | -0.019<br>(.078) | -0.550***<br>(.202) |
| FACE | 0.452***<br>(.109) | 0.407***<br>(.110) | -- | 0.408***<br>(.110) |
| BLANK | .923***<br>(.165) | | -- | |
| SPRING | -- | 0.384***<br>(.117) | 0.383***<br>(.118) | 0.406***<br>(.118) |
| BLANK/ SPRING | -1.106**<br>(.237) | | -- | |
| Pair DEVIOUS | -- | | .470***<br>(.163) | |
| Pair HAPPY | -- | | 0.367**<br>(.159) | |
| Pair OVAL | -- | | -0.207<br>(.207) | |
| Pair RECTANGLE | | | -0.040<br>(.210) | |
| GAIN | | | | 0.050**<br>(.022) |
| | n=727<br>LL=-463.12 | n=578<br>LL=-376.81 | n=578<br>LL=-377.41 | n=578<br>LL=-374.20 |

***p<.01<br>
**p<.05<br>
*p<.1

## Table 4
## Probit Estimates for Choosing "Trust" Move or Reciprocated Trust (Standard Errors and p-values in Parentheses)

<Note: These are for Prediction 3 -- "Trust" is a move down at node 3 while reciprocity is a move down at node 4 -- each is a different actor's move, as a function of the other's icon that is observed>

| | Trust | Reciprocity |
|---|---|---|
| INTERCEPT | .566<br>(.502, p=.260) | -.566<br>(.502, p=.260) |
| DEVIOUS | .068<br>(.590, p=.909) | .776<br>(.621, p=.211) |
| HAPPY | -.343<br>(.589. p=.560) | .997<br>(.626, p=.112) |
| RECTANGLE | -.566<br>(.619, p=.361) | .566<br>(.803, p=.481) |
| | n=55<br>LL=-34.97 | n=35<br>LL=-22.74 |

***p<.01
**p<.05
*p<.1

**Table 5**
**Probit Estimates for "Trusting" Given Past History**
**(Standard Errors in Parentheses)**

| | Model 1 | Model 2 |
|---|---|---|
| INTERCEPT | -.365<br>(.225) | -.322<br>(.256) |
| DEVIOUS | .501***<br>(.168) | .519***<br>(.179) |
| HAPPY | .309*<br>(.169) | .361**<br>(.179) |
| RECTANGLE | -.077<br>(.176) | -.049<br>(.182) |
| GAIN | .045**<br>(.021) | .039*<br>(.023) |
| BETRAYED | -- | .003<br>(.002) |
| BETRAYER | -- | -.014***<br>(.002) |
| EARNINGS | -- | .028<br>(.017) |
| | n=578<br>llr=-379.09 | n=578<br>LL=-350.46 |

***p<.01
**p<.05
*p<.1

### Table 6
### Reciprocated Trust
### (Standard Errors in Parentheses)

| | Dependent Variable | | |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| INTERCEPT | -.408*<br>(.226) | .621**<br>(.293) | .698**<br>(.295) |
| DEVIOUS | .190<br>(.200) | .029<br>(.215) | -.004<br>(.218) |
| HAPPY | .450**<br>(.193) | .337*<br>(.206) | .329<br>(.208) |
| RECTANGLE | .114<br>(.264) | .023<br>(.286) | .026<br>(.287) |
| TEMPTATION | -.060***<br>(.014) | -.071***<br>(.015) | -.077***<br>(.016) |
| BETRAYED | -- | .004<br>(.003) | .007<br>(.003) |
| BETRAYER | -- | -.021***<br>(.003) | -.019***<br>(.003) |
| EARNINGS | -- | .049**<br>(.021) | .052**<br>(.021) |
| LAST EXPERIENCE | -- | -- | -.453**<br>(.183) |
| | n=342<br>LL=-223.09 | n=342<br>LL=-191.64 | n=342<br>LL=-188.55 |

***p<.01
**p<.05
*p<.1

Figure 1: Game A

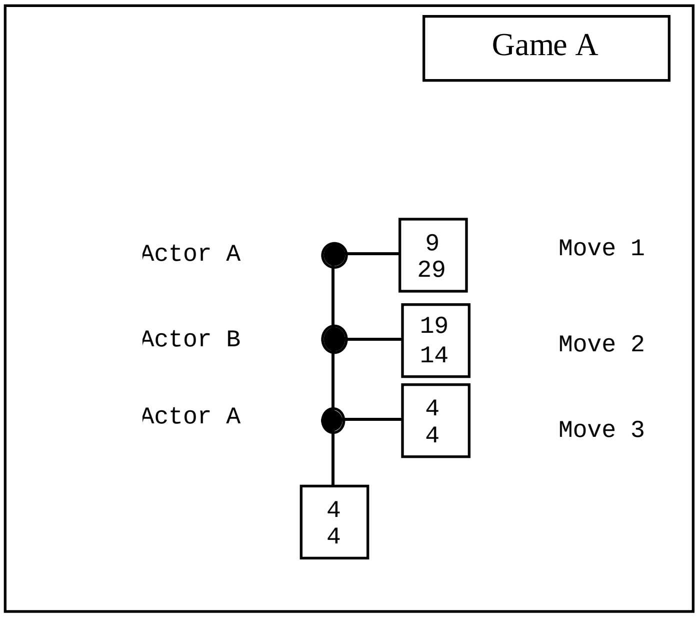


Figure 2: Game B

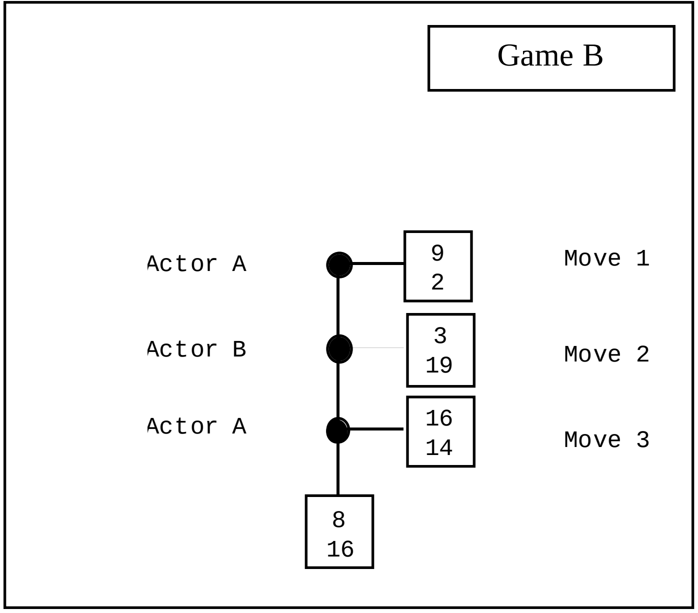


Figure 3

Icons Used in Survey

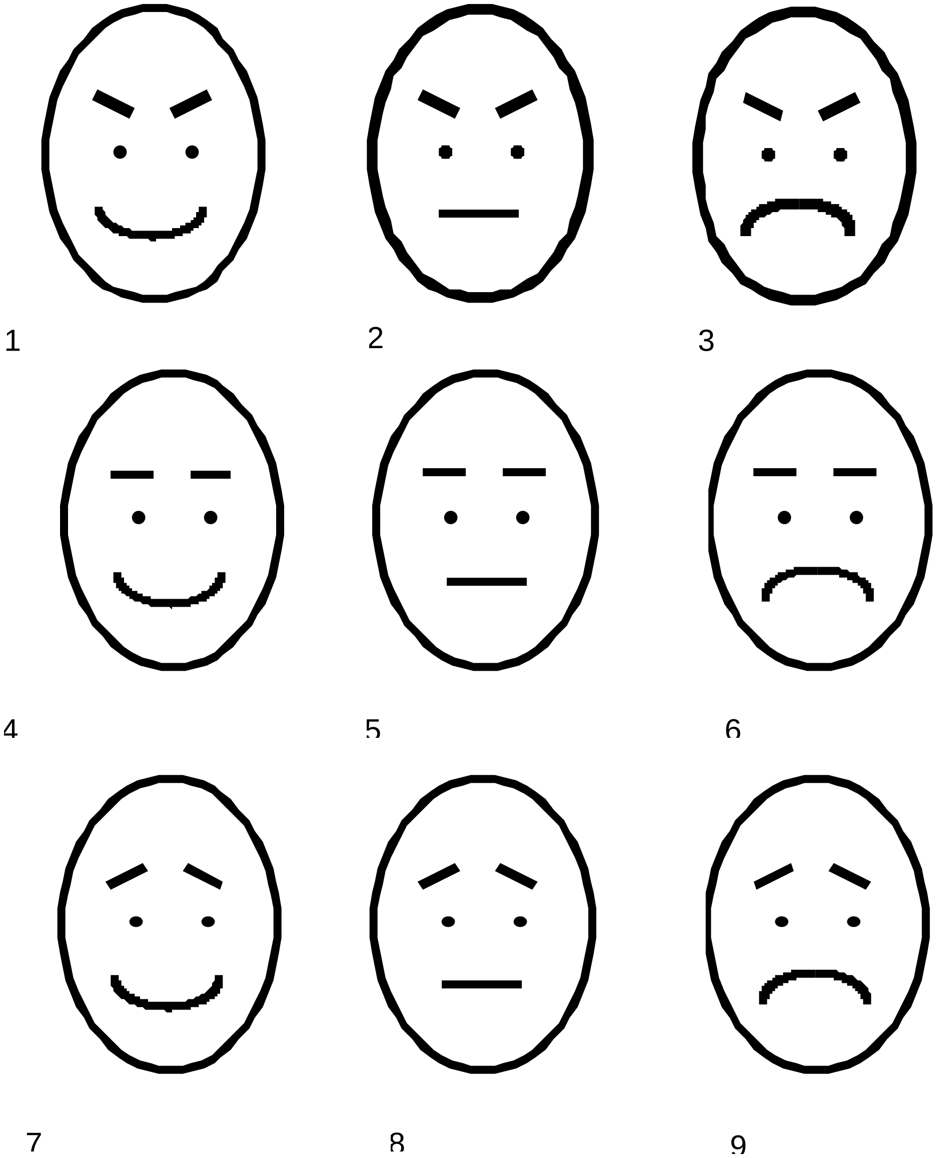


Figure 4
Mean Ratings of Icons on Semantic Differential Items

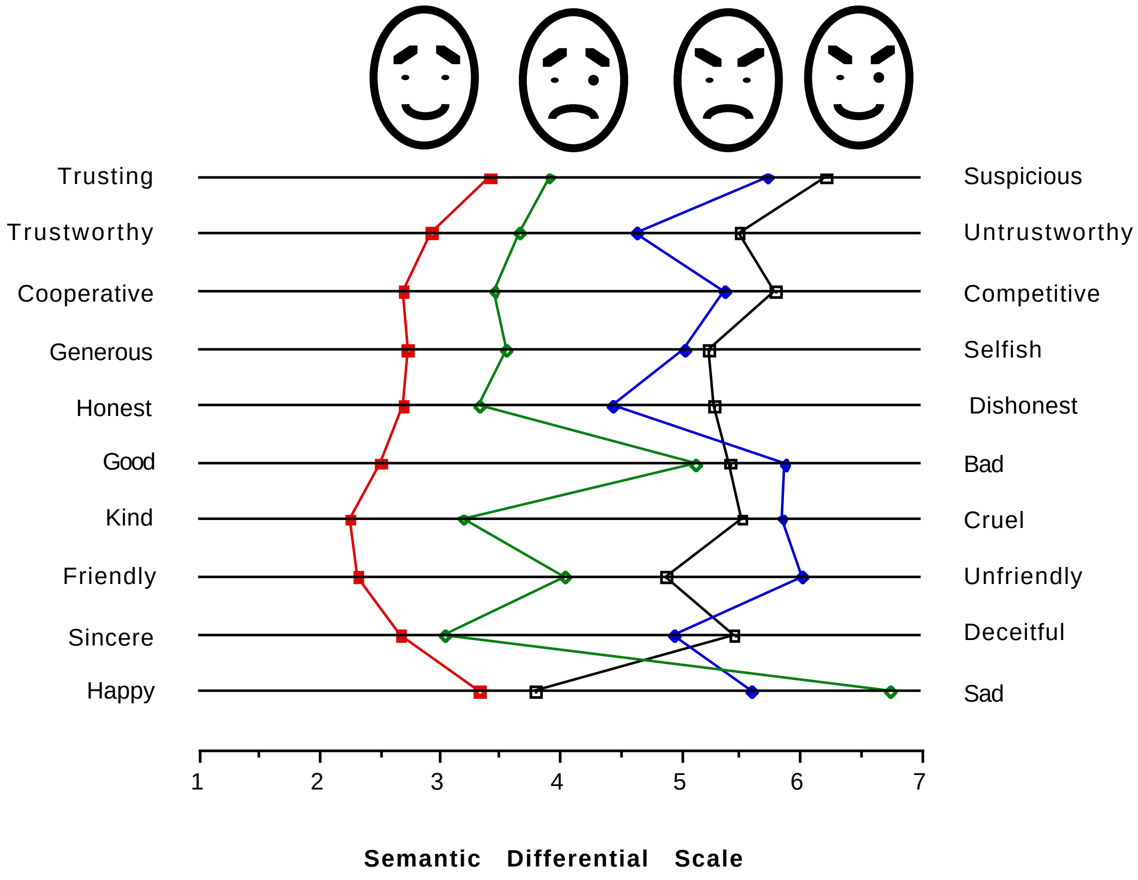


Figure 5
A Subset of Games and Parameters used in Series A and B Experiments

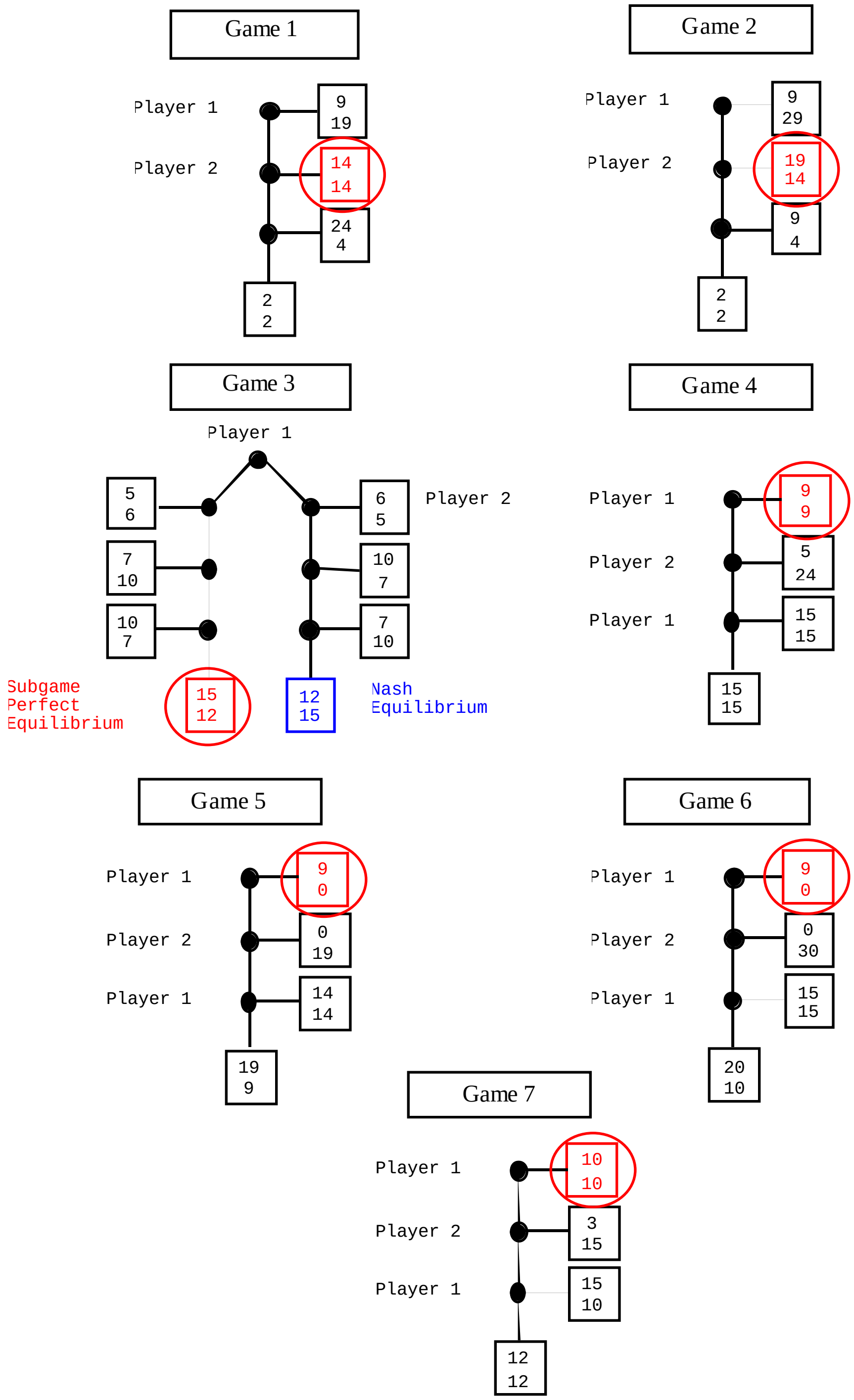


Figure 6
Icons Used in Series A and B Experiments

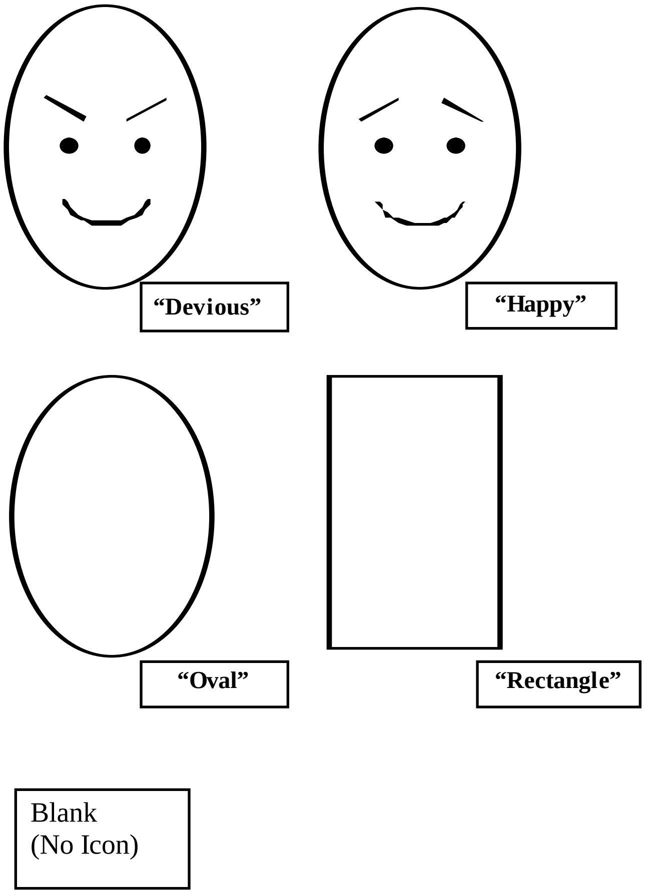


Figure 7
Aggregate Outcomes for Simple Bargaining Games

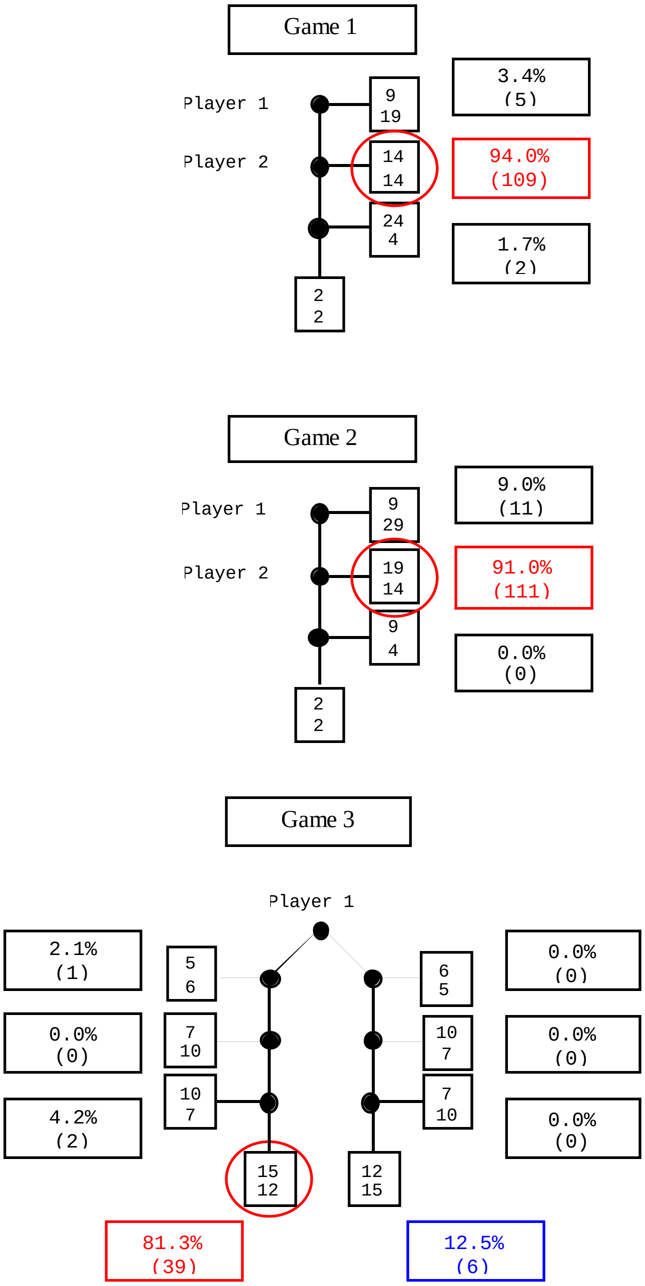
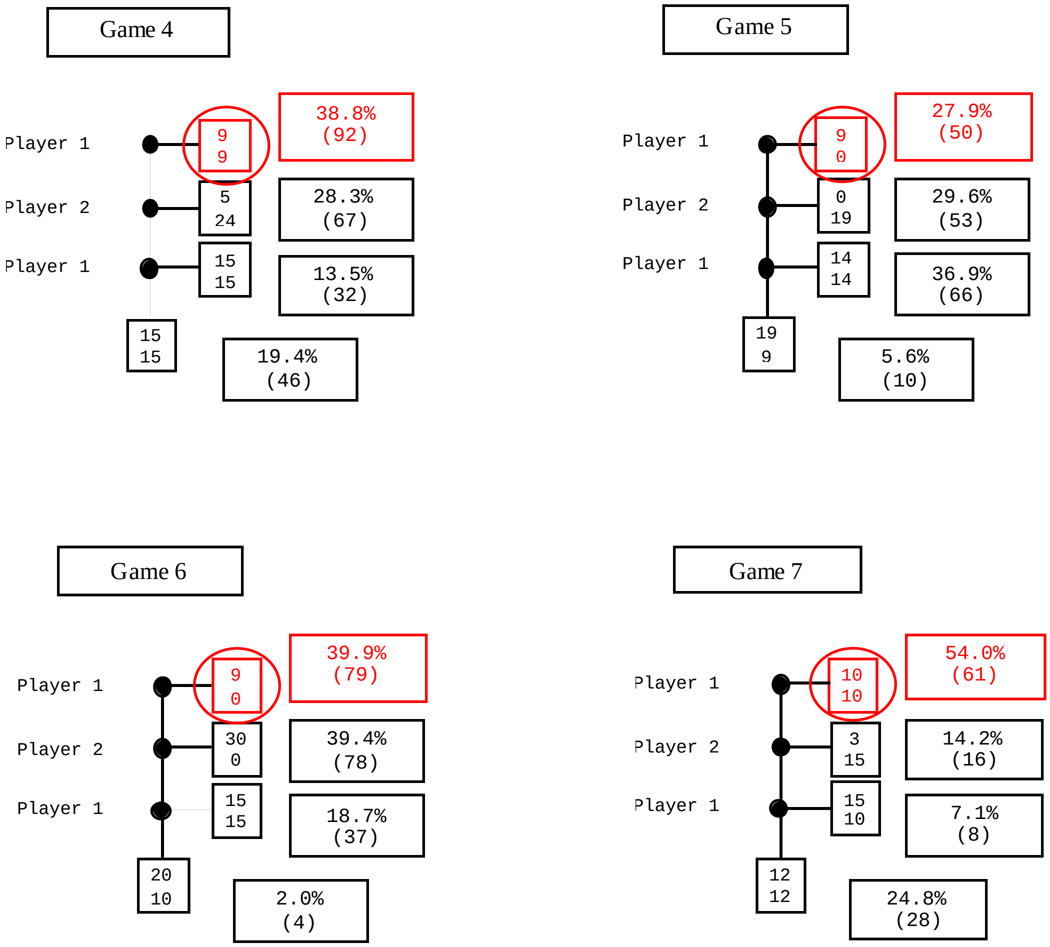
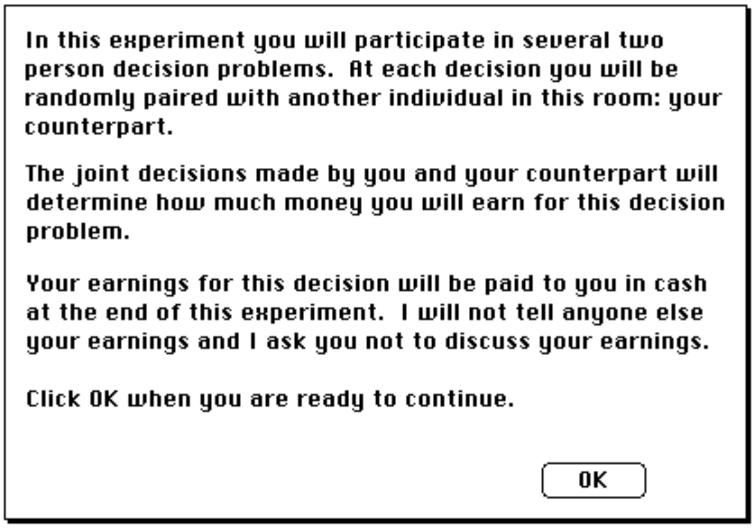


Figure 8
Aggregate Outcomes for Trust Games



# Appendix 1
# Instruction Set

Screen 1



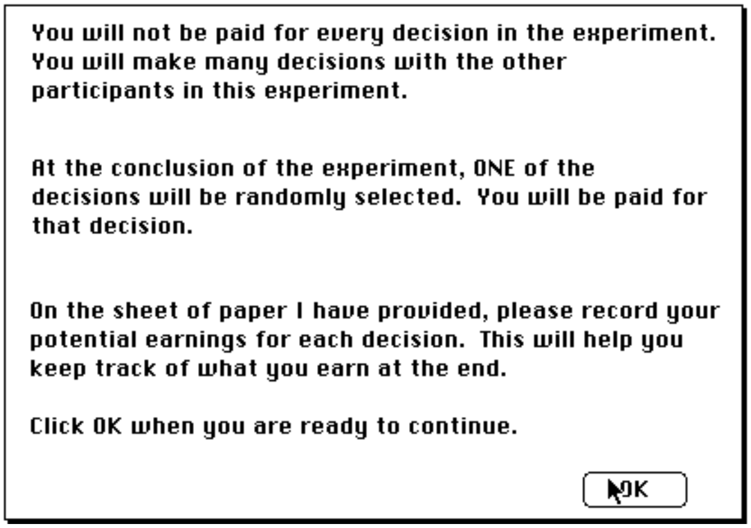
In this experiment you will participate in several two person decision problems. At each decision you will be randomly paired with another individual in this room: your counterpart.

The joint decisions made by you and your counterpart will determine how much money you will earn for this decision problem.

Your earnings for this decision will be paid to you in cash at the end of this experiment. I will not tell anyone else your earnings and I ask you not to discuss your earnings.

Click OK when you are ready to continue.

Screen 2



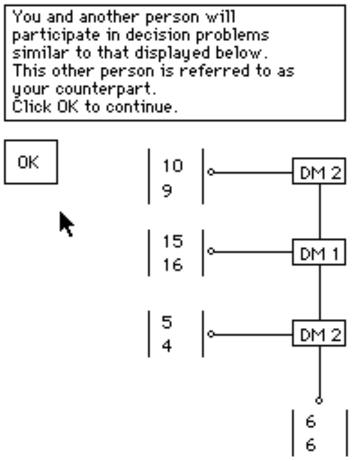
You will not be paid for every decision in the experiment. You will make many decisions with the other participants in this experiment.

At the conclusion of the experiment, ONE of the decisions will be randomly selected. You will be paid for that decision.

On the sheet of paper I have provided, please record your potential earnings for each decision. This will help you keep track of what you earn at the end.

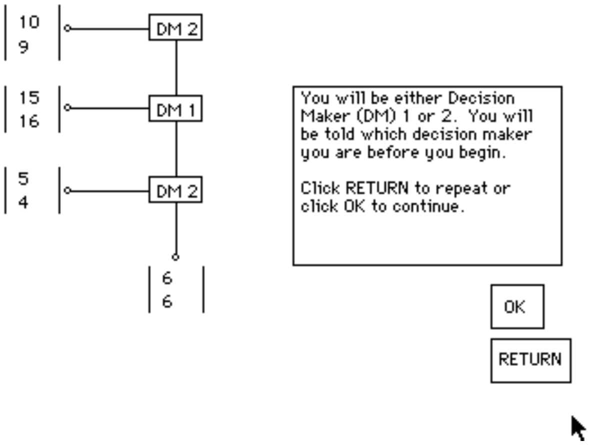
Click OK when you are ready to continue.

Screen 3



You and another person will participate in decision problems similar to that displayed below. This other person is referred to as your counterpart.
Click OK to continue.

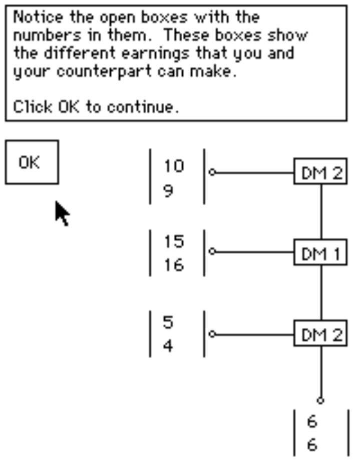
Screen 4



You will be either Decision Maker (DM) 1 or 2. You will be told which decision maker you are before you begin.

Click RETURN to repeat or click OK to continue.

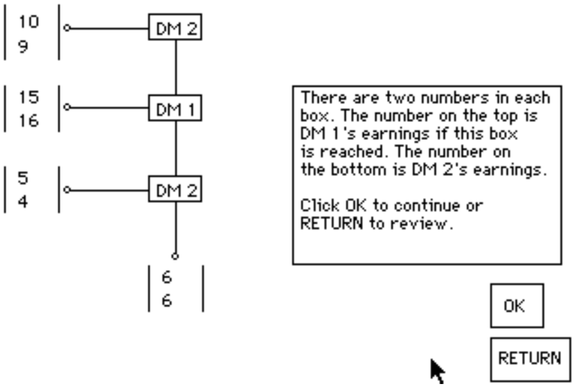
Screen 5



Notice the open boxes with the numbers in them. These boxes show the different earnings that you and your counterpart can make.

Click OK to continue.

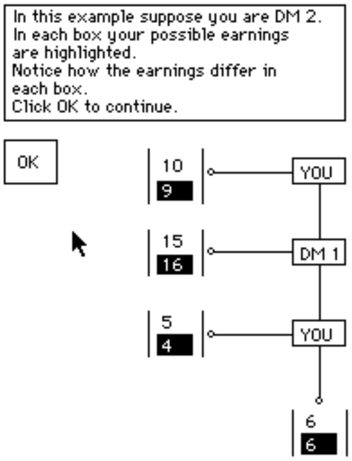
Screen 6



There are two numbers in each box. The number on the top is DM 1's earnings if this box is reached. The number on the bottom is DM 2's earnings.

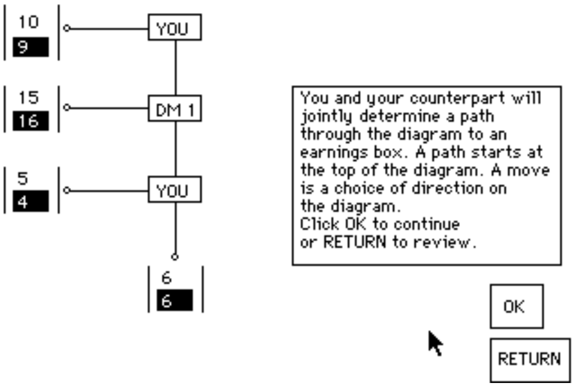
Click OK to continue or RETURN to review.

Screen 7



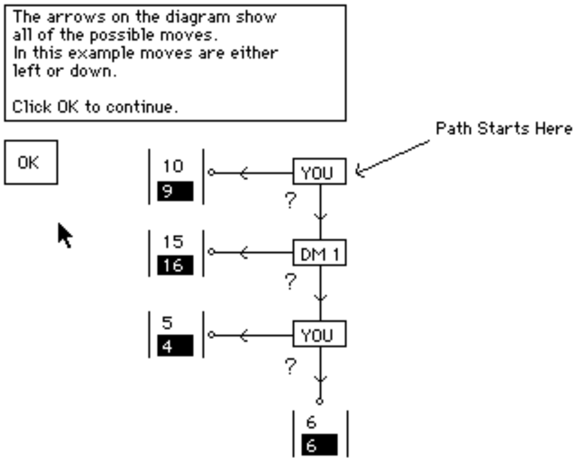
In this example suppose you are DM 2.
In each box your possible earnings
are highlighted.
Notice how the earnings differ in
each box.
Click OK to continue.

Screen 8



You and your counterpart will
jointly determine a path
through the diagram to an
earnings box. A path starts at
the top of the diagram. A move
is a choice of direction on
the diagram.
Click OK to continue
or RETURN to review.

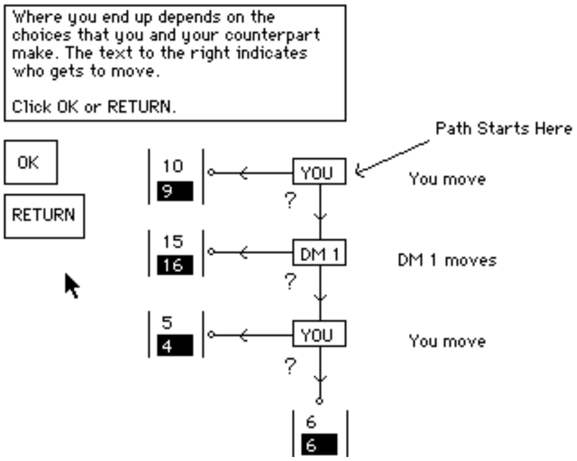
Screen 9



The arrows on the diagram show all of the possible moves.
In this example moves are either left or down.

Click OK to continue.

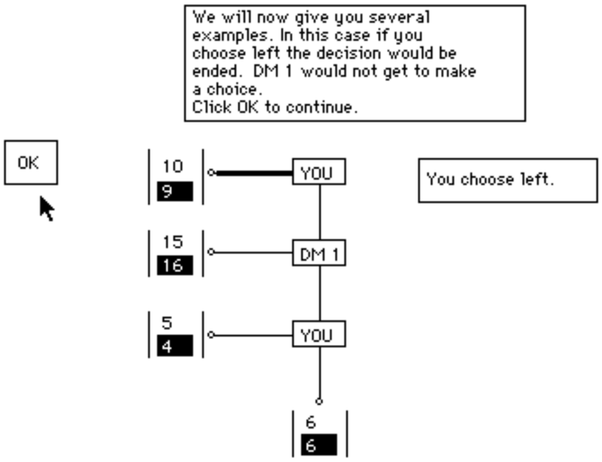
Screen 10



Where you end up depends on the choices that you and your counterpart make. The text to the right indicates who gets to move.

Click OK or RETURN.

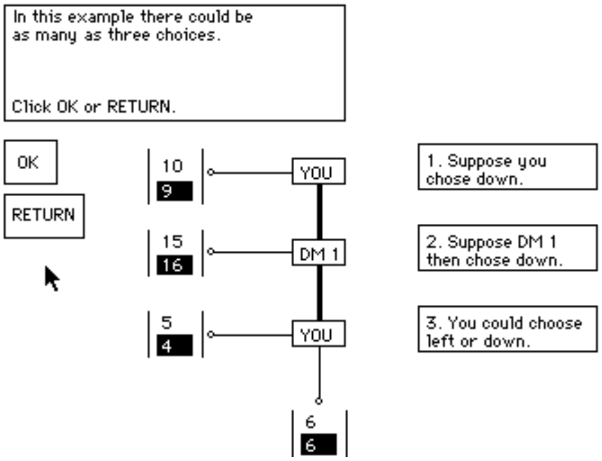
Screen 11



We will now give you several examples. In this case if you choose left the decision would be ended. DM 1 would not get to make a choice.
Click OK to continue.

You choose left.

Screen 12



In this example there could be as many as three choices.

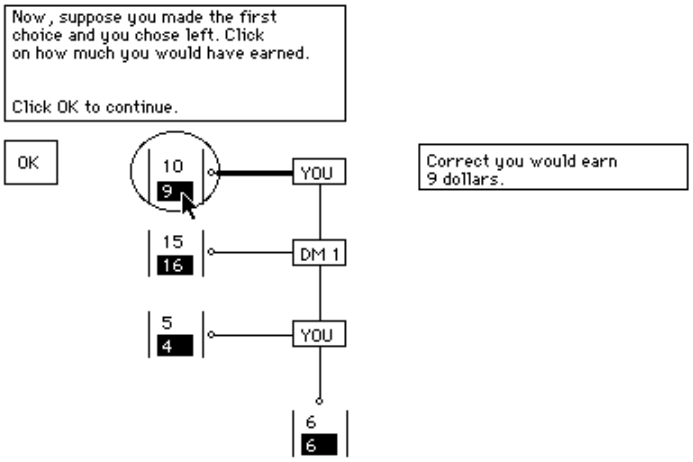
Click OK or RETURN.

1. Suppose you chose down.

2. Suppose DM 1 then chose down.

3. You could choose left or down.

Screen 13

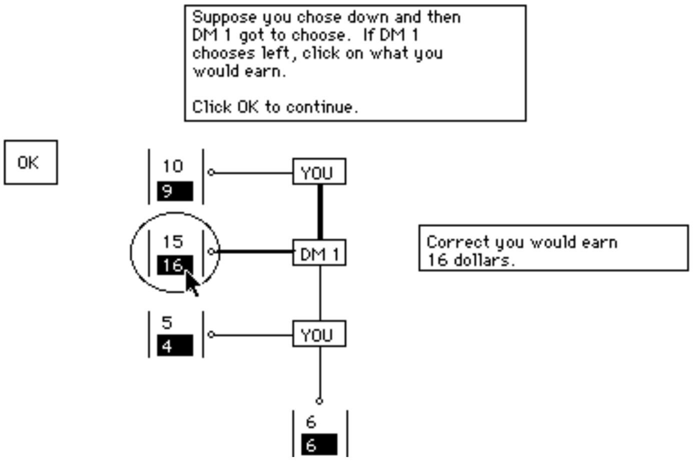


Now, suppose you made the first choice and you chose left. Click on how much you would have earned.

Click OK to continue.

Correct you would earn 9 dollars.

Screen 14

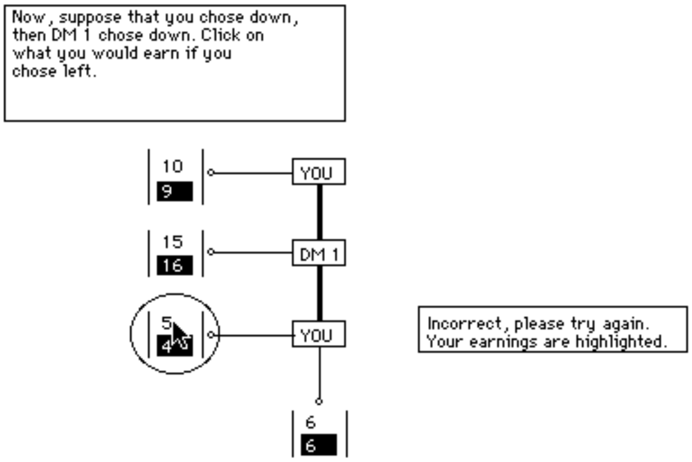


Suppose you chose down and then DM 1 got to choose. If DM 1 chooses left, click on what you would earn.

Click OK to continue.

Correct you would earn 16 dollars.

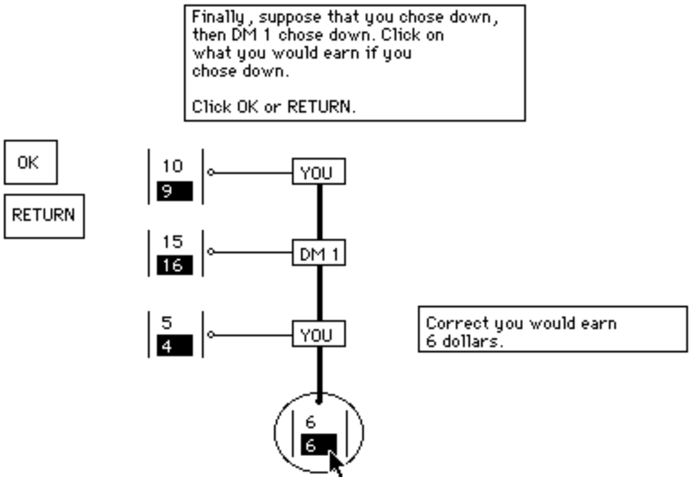
Screen 15



Now, suppose that you chose down, then DM 1 chose down. Click on what you would earn if you chose left.

Incorrect, please try again. Your earnings are highlighted.

Screen 16

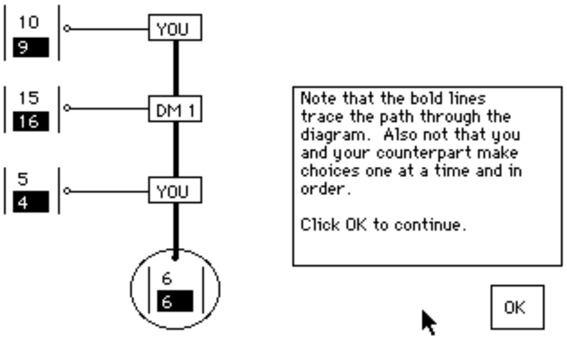


Finally, suppose that you chose down, then DM 1 chose down. Click on what you would earn if you chose down.

Click OK or RETURN.

Correct you would earn 6 dollars.

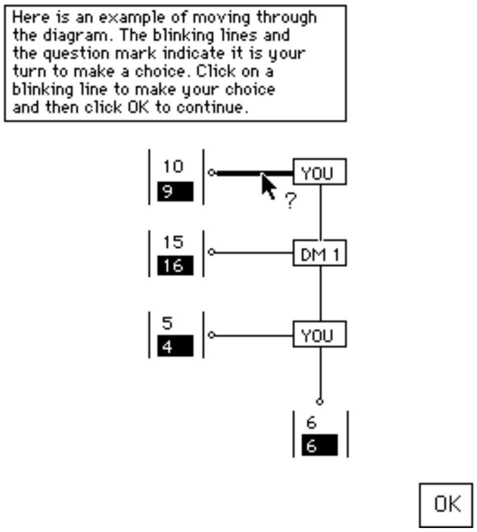
Screen 17



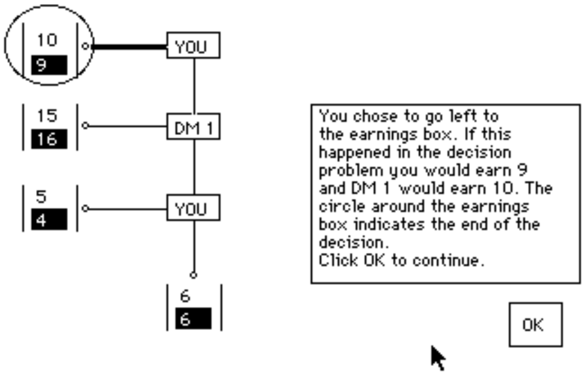
Note that the bold lines trace the path through the diagram. Also not that you and your counterpart make choices one at a time and in order.

Click OK to continue.

Screen 18

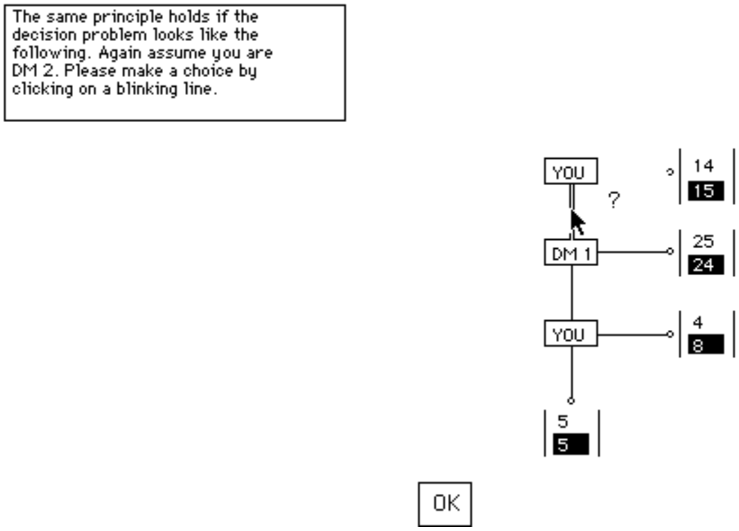

Here is an example of moving through the diagram. The blinking lines and the question mark indicate it is your turn to make a choice. Click on a blinking line to make your choice and then click OK to continue.

Screen 19



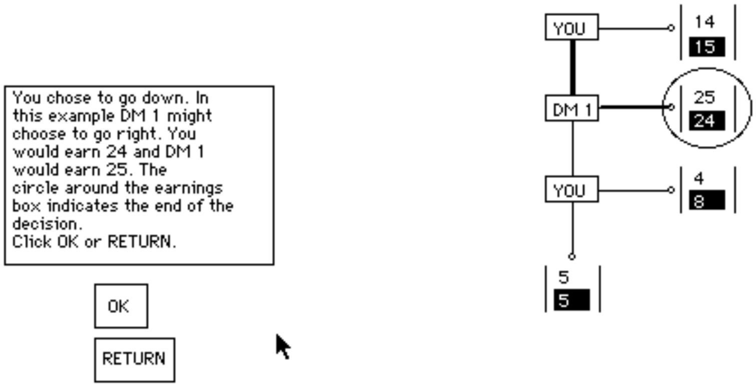
You chose to go left to the earnings box. If this happened in the decision problem you would earn 9 and DM 1 would earn 10. The circle around the earnings box indicates the end of the decision.
Click OK to continue.

Screen 20



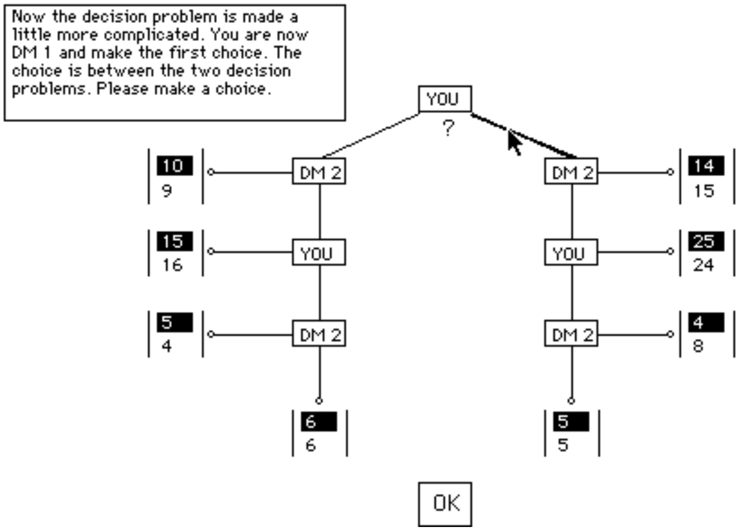
The same principle holds if the decision problem looks like the following. Again assume you are DM 2. Please make a choice by clicking on a blinking line.

Screen 21



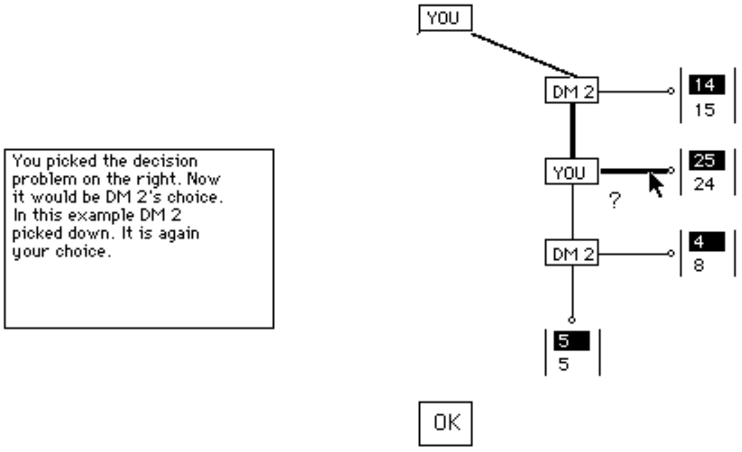
You chose to go down. In this example DM 1 might choose to go right. You would earn 24 and DM 1 would earn 25. The circle around the earnings box indicates the end of the decision.
Click OK or RETURN.

Screen 22



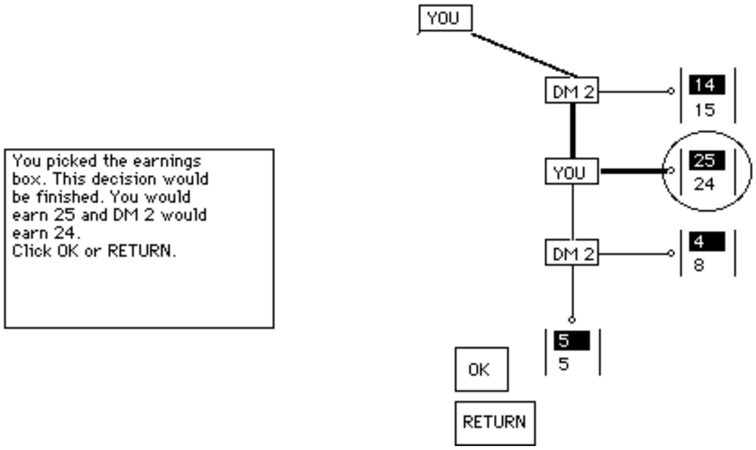
Now the decision problem is made a little more complicated. You are now DM 1 and make the first choice. The choice is between the two decision problems. Please make a choice.

Screen 23



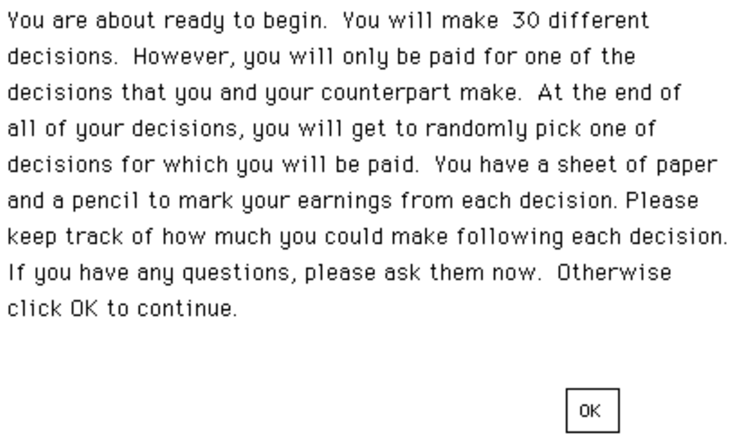
You picked the decision problem on the right. Now it would be DM 2's choice. In this example DM 2 picked down. It is again your choice.

Screen 24



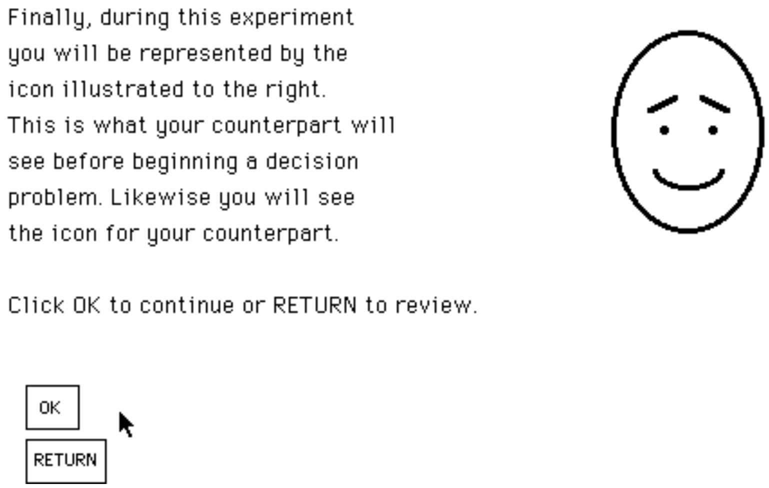
You picked the earnings box. This decision would be finished. You would earn 25 and DM 2 would earn 24.
Click OK or RETURN.

Screen 25



You are about ready to begin. You will make 30 different decisions. However, you will only be paid for one of the decisions that you and your counterpart make. At the end of all of your decisions, you will get to randomly pick one of decisions for which you will be paid. You have a sheet of paper and a pencil to mark your earnings from each decision. Please keep track of how much you could make following each decision. If you have any questions, please ask them now. Otherwise click OK to continue.

Screen 26



Finally, during this experiment you will be represented by the icon illustrated to the right. This is what your counterpart will see before beginning a decision problem. Likewise you will see the icon for your counterpart.

Click OK to continue or RETURN to review.